Review Article

A narrative review of foundation models for medical image segmentation: zero-shot performance evaluation on diverse modalities

Seungha Noh, Byoung-Dai Lee

Department of Computer Science, Graduate School, Kyonggi University, Suwon, Republic of Korea

Contributions: (I) Conception and design: BD Lee; (II) Administrative support: BD Lee; (III) Provision of study materials or patients: Both authors; (IV) Collection and assembly of data: Both authors; (V) Data analysis and interpretation: Both authors; (VI) Manuscript writing: Both authors; (VII) Final approval of manuscript: Both authors.

Correspondence to: Byoung-Dai Lee, PhD. Department of Computer Science, Graduate School, Kyonggi University, 154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16227, Republic of Korea. Email: blee@kgu.ac.kr.

Background and Objective: Foundation models are deep learning models pretrained on extensive datasets, equipped with the ability to adapt to a variety of downstream tasks. Recently, they have gained prominence across various domains, including medical imaging. These models exhibit remarkable contextual understanding and generalization capabilities, spurring active research in healthcare to develop versatile artificial intelligence solutions for real-world clinical environments. Inspired by this, this study offers a comprehensive review of foundation models in medical image segmentation (MIS), evaluates their zero-shot performance on diverse datasets, and assesses their practical applicability in clinical settings.

Methods: A total of 63 studies on foundation models for MIS were systematically reviewed, utilizing platforms such as arXiv, ResearchGate, Google Scholar, Semantic Scholar, and PubMed. Additionally, we curated 31 unseen medical image datasets from The Cancer Imaging Archive (TCIA), Kaggle, Zenodo, Institute of Electrical and Electronics Engineers (IEEE) DataPort, and Grand Challenge to evaluate the zero-shot performance of six foundation models. Performance analysis was conducted from various perspectives, including modality and anatomical structure.

Key Content and Findings: Foundation models were categorized based on a taxonomy that incorporates criteria such as data dimensions, modality coverage, prompt type, and training strategy. Furthermore, the zero-shot evaluation revealed key insights into their strengths and limitations across diverse imaging modalities. This analysis underscores the potential of these models in MIS while highlighting areas for improvement to optimize real-world applications.

Conclusions: Our findings provide a valuable resource for understanding the role of foundation models in MIS. By identifying their capabilities and limitations, this review lays the groundwork for advancing their practical deployment in clinical environments, supporting further innovation in medical image analysis.

Keywords: Deep learning; foundation model; medical imaging segmentation; zero-shot performance

Submitted Dec 12, 2024. Accepted for publication Mar 27, 2025. Published online Jun 03, 2025.
doi: 10.21037/qims-2024-2826
View this article at: https://dx.doi.org/10.21037/qims-2024-2826

Introduction

Medical image segmentation (MIS) aims to accurately identify specific anatomical structures such as organs, lesions, and tissues. It is widely applied across various medical domains, including radiology and pathology, for operations such as surgical planning, as well as in the diagnosis and prognosis of diseases (1,2). Consequently, research on deep learning-based methods to ensure consistent and precise MIS has been actively conducted.

Early MIS approaches were relied on traditional image processing algorithms such as thresholding, region growing, edge detection, and the Markov random fields (3). These methods have limitations as they are labor-intensive and sensitive to variations. With the advancement of deep learning techniques, models based on convolutional neural networks (CNNs) (4) began to emerge. Particularly, U-Net (5) is widely recognized as a benchmark model for MIS due to its impressive performance and straightforward architecture. Building on this, numerous extensions such as ResU-Net (6), U-Net++ (7), and three-dimensional (3D) U-Net (8) have been developed. However, CNNs have inherent limitations in fully capturing long-range dependencies due to their fixed receptive fields.

Transformer (9) was introduced in MIS due to its ability to capture global contextual information, overcoming the inherent limitations of CNNs regarding restricted receptive fields. Building on this, extensive research has been conducted to combine the advantages of CNNs and Transformers (10-14). These efforts have led to models that achieve high performance across various medical imaging tasks. However, existing MIS models tend to be optimized for specific tasks, which may lead to performance degradation when dealing with new tasks or different types of data. The limited scope of applicability poses a significant obstacle to the implementation of these models in real-world clinical environments.

Recently, large-scale foundation models have significantly advanced the field of artificial intelligence, demonstrating remarkable zero-shot and few-shot performance across various downstream tasks. Particularly, the segment anything model (SAM) (15), a foundation model for natural image segmentation, has not only demonstrated robust performance across various segmentation tasks but has also set a new standard in image segmentation by utilizing prompt-based interaction. Inspired by these, numerous studies have attempted to extend SAM to the medical domain. However, the significant domain gap between medical- and natural-images presents challenges in directly applying foundation models to medical domains (16-21).

The challenges can be attributed to three main factors. First, medical images differ significantly from natural images in appearance, with distinct differences in visual aspects such as color, brightness, and contrast. Medical images are categorized into different modalities, including X-ray, computed tomography (CT), and ultrasound, based on the specific imaging technique and equipment utilized, each having its own distinct features. Furthermore, 3D volumetric data from modalities like CT and magnetic resonance imaging (MRI) provide information in an additional dimension compared to two-dimensional (2D) images, which must be considered. Second, the boundaries of target structures in medical images are often blurred, making it common for tissues and organs to exhibit indistinct edges. While experienced medical experts can identify these subtle boundaries through their deep understanding of anatomy, deep learning models trained primarily on natural images struggle to recognize these fine distinctions. Lastly, large medical datasets are needed to address these challenges. Due to the diverse and complex characteristics of medical images, incorporating various modalities and conditions is crucial for the deep learning models to learn effectively.

Numerous studies have investigated the adaptation of vision foundation models to the medical domain to overcome existing constraints (22-25). However, their applicability in real clinical settings remains underexplored, with insufficient validation of performance on unseen datasets. To address this gap, we provide a comprehensive review of research on foundation models in MIS, analyzing them from various perspectives to offer an in-depth understanding of their capabilities and limitations. Additionally, we assessed the zero-shot performance of six representative models across 31 unseen datasets, including diverse modalities and anatomical structures, to evaluate their robustness and practical utility. This study offers valuable insights into current methodologies and highlights potential directions for future research in the development and application of foundation models. We present this article in accordance with the Narrative Review reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-2024-2826/rc).

Methods

In this survey, we conducted a systematic search for

Table 1 The search strategy summary

Items	Specification
Date of search	4 September 2024
Database and other sources searched	arXiv, ResearchGate, Google Scholar, Semantic Scholar, PubMed
Search terms used	'Medical', 'Healthcare', 'Foundation', 'Universal', 'General', 'Segmentation'
Timeframe	2023.02–2024.09
Inclusion and exclusion criteria	Inclusion: medical image segmentation studies focusing on generalization, published in English
	Exclusion: studies not related to segmentation networks, studies outside the medical domain, and conference abstracts only
Selection process	Authors S.N. and B.D.L. performed joint selection

the recent studies on foundation models in MIS using academic databases and platforms such as arXiv (26), ResearchGate (27), Google Scholar (28), Semantic Scholar (29), and PubMed (30). The search was conducted on 4 September 2024. To collect pertinent papers, we adopted a two-step selection process. First, we implemented a keyword-based search strategy by categorizing keywords into three categories: Domain, Feature, and Task. For the Domain category, we selected keywords such as "Medical" and "Healthcare". For the Feature category, we focused on terms like "General", "Foundation", and "Universal". Finally, the Task category was defined by the keyword "Segmentation". We constructed a search query by combining keywords using Boolean operators. Specifically, keywords within each category were linked with the "OR" operator, while the different categories were connected using the "AND" operator. The final search query was structured as follows: {("Medical" OR "Healthcare") AND ("General" OR "Foundation" OR "Universal") AND ("Segmentation")}. Second, we evaluated the chosen papers and excluded any that were irrelevant to our study. In the initial screening stage, we assessed each paper's relevance by reviewing its title and abstract, followed by a full-text review for final categorization according to the taxonomy defined in this study. We excluded papers that did not directly address segmentation tasks, studies that were outside the medical domain, and those available only as conference abstracts. This process began with an initial identification of 143 papers through the Boolean search query. After evaluating their relevance based on predefined inclusion and exclusion criteria, we finalized a total of 63 papers for analysis. Details of the research methodology and paper selection criteria are summarized in *Table 1*.

Architecture of the SAM

Recently, with the emergence of powerful language foundation models such as GPT-4 (31), deep learning vision models pretrained on large-scale datasets have paved the way for new research directions (15,32-36). These models demonstrate deep representational capacity and exceptional generalization performance, consistently excelling in various downstream tasks. Specifically, there is significant potential to accelerate the development of accurate and robust models in medical image analysis (37,38).

Among these, the SAM stands out as the first prompt-based image segmentation foundation model. By utilizing transformer-based architecture, SAM has demonstrated exceptional generalization performance through large-scale training on the SA-1B dataset, which comprises 11M diverse general images. Its robustness and impressive results in zero-shot transfer experiments have inspired the development of various other foundation models. As shown in *Figure 1*, SAM consists of three main components: an image encoder, a prompt encoder, and a mask decoder. The image encoder utilizes the Vision Transformer (ViT) (10) architecture, pre-trained with the masked autoencoder (MAE) (40) method, to convert high-resolution images into image embeddings. Next, the prompt encoder represents various prompts (e.g., point, box) as embeddings. Finally, the mask decoder integrates the image embeddings and prompt embeddings via a cross-attention mechanism to predict the mask for the region indicated by the prompt. The intersection over union (IoU) head then evaluates the IoU score. This design enables SAM to provide an interactive, user-friendly, and flexible segmentation framework.

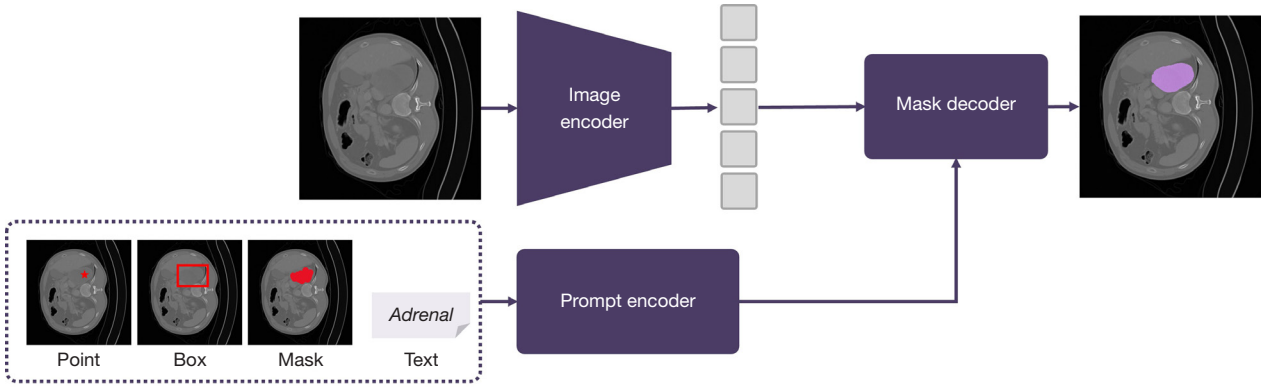


Figure 1 The architecture of SAM, comprising three components: image encoder, prompt encoder, and mask decoder. Medical image was obtained from the publicly available Adrenal-ACC-Ki67-Seg dataset (39). SAM, segment anything model.

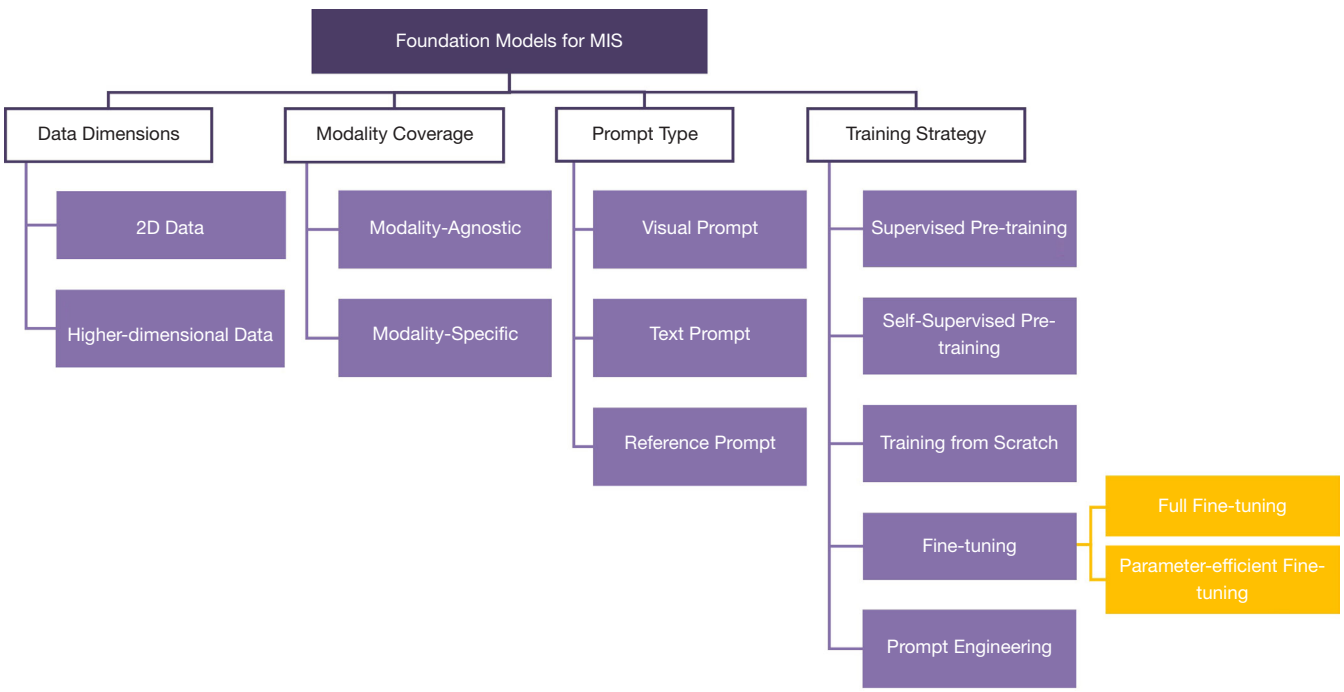


Figure 2 The taxonomy consists of four distinct criteria for categorizing foundation models for MIS: (I) data dimensions; (II) modality coverage; (III) prompt type; and (IV) training strategy. 2D, two-dimensional; MIS, medical image segmentation.

Foundation models for MIS

This section provides an in-depth introduction to foundation models for MIS, organized using a taxonomy based on four criteria: (I) data dimensions; (II) modality coverage; (III) prompt type; and (IV) training strategy, as illustrated in *Figure 2*. This taxonomy is provided to systematically analyze and present the models, and a summary of all foundation models discussed in this paper is provided in *Table 2*.

Data dimensions

Medical imaging encompasses a variety of dimensions, ranging from 2D images—such as X-rays, mammography,

Table 2 The summary of foundation models in medical image segmentation

Method	Dimension	Modalities	Training	Year and month	Code
MedFormer (41)	3D	MR	Scratch	2023.02	✓
MedCLIP-SAM (42)	2D	Multi	PE	2023.03	✓
Med-SA (43)	2D	Multi	FT, PE	2023.04	✓
SkinSAM (44)	2D	Dermoscopy	FT	2023.04	
STU-Net (45)	3D	CT	SP, FT	2023.04	✓
UniverSeg (46)	2D	Multi	SP, PE	2023.04	✓
Polyp-SAM (47)	2D	Endoscopy	FT	2023.04	✓
SAMed (48)	2D	CT	FT	2023.04	✓
MedSAM (22)	2D	Multi	FT	2023.04	✓
SAM-LST (49)	2D	CT	FT	2023.06	✓
3DSAM-Adapter (50)	3D	Multi	FT	2023.06	✓
MIS-FM (51)	3D	CT	SSP, FT	2023.06	✓
AutoSAM (52)	2D	Multi	PE	2023.06	✓
SAM-Path (53)	2D	Pathology	PE	2023.07	✓
Shi *et al.* (54)	3D	MR	FT	2023.07	
tUbe net (55)	3D	Multi	Scratch	2023.07	✓
SurgicalSAM (56)	2D	Endoscopy	PE	2023.08	✓
Polyp-SAM++ (57)	2D	Endoscopy	PE	2023.08	✓
AdaptiveSAM (58)	2D	Multi	FT	2023.08	✓
SAM-Med2D (23)	2D	Multi	FT	2023.08	✓
SAM3D (59)	3D	Multi	FT	2023.09	✓
SAM-OCTA (60)	2D	OCT	FT	2023.09	✓
Anand *et al.* (61)	2D	Multi	PE	2023.09	
MediViSTA-SAM (62)	2D	US	FT	2023.09	✓
MA-SAM (63)	3D	Multi	FT	2023.09	✓
SonoSAMTrack (64)	2D	US	FT	2023.10	
WSI-SAM (65)	2D	Pathology	FT	2023.10	✓
Pandey *et al.* (66)	2D	Multi	PE	2023.10	
SAMPOT (67)	2D	X-ray	PE	2023.10	
SonoSAM (68)	2D	US	FT	2023.10	
SAM-Med3D (24)	3D	Multi	Scratch	2023.10	✓
Eviprompt (69)	2D	Multi	PE	2023.11	✓
MedLSAM (70)	3D	CT	PE	2023.11	✓
CellSAM (71)	2D	Multi	PE	2023.11	✓
SegVol (72)	3D	CT	SSP, FT, PE	2023.11	✓

Table 2 (*continued*)

Table 2 (*continued*)

Method	Dimension	Modalities	Training	Year and month	Code
GMISeg (73)	2D	Multi	PE	2023.12	
SAM-CLNet (74)	2D	Endoscopy	Scratch	2023.12	
SP-SAM (75)	2D	Endoscopy	FT, PE	2023.12	✓
SAT (25)	3D	Multi	PE	2023.12	✓
AFTer-SAM (76)	3D	CT	FT	2024.01	
PUNETR (77)	2D	CT	SSP, PE	2024.01	✓
SegmentAnyBone (78)	3D	MR	FT, PE	2024.01	✓
UN-SAM (79)	2D	Microscopy	FT, PE	2024.02	✓
APPLE (80)	2D	Multi	FT	2024.03	
FluoroSAM (81)	2D	X-ray	FT, PE	2024.03	✓
ProMISe (82)	2D	Multi	PE	2024.03	✓
SAIM (83)	2D	US	FT	2024.03	
Zhou and Yu (84)	2D	MR	FT	2024.03	
One-Prompt Segmentation (85)	2D	Multi	PE	2024.04	✓
Indelman *et al.* (86)	2D	US	–	2024.04	
MAFUnet (87)	3D	MR	SP, FT	2024.05	✓
GSAM+ Cutie (88)	2D	Endoscopy	FT, PE	2024.06	✓
Hermes (89)	2D	Multi	Scratch	2024.06	✓
WSPolyp-SAM (90)	2D	Endoscopy	FT	2024.06	
MoME (91)	3D	MR	FT	2024.07	✓
ESP-MedSAM (92)	2D	Multi	PE	2024.07	✓
DrSAM (93)	2D	Multi	FT, PE	2024.07	✓
CC-SAM (94)	2D	US	FT, PE	2024.07	
SAMUS (95)	2D	US	FT, PE	2024.07	✓
DeSAM (96)	2D	Multi	PE	2024.07	✓
SAM-UNet (97)	2D	Multi	FT	2024.08	✓
BrainSegFounder (98)	3D	MR	SSP, FT	2024.08	✓
SimSAM (18)	2D, 3D	Multi	–	2024.08	✓

2D, two-dimensional; 3D, three-dimensional; CT, computed tomography; FT, fine-tuning; MR, magnetic resonance; OCT, optical coherence tomography; PE, prompt engineering; Scratch, training from scratch; SP, supervised pre-training; SSP, self-supervised pre-training; US, ultrasound.

and endoscopy—to higher-dimensional data, including CT, MRI, positron emission tomography (PET), and functional MRI (fMRI). Consequently, model design must consider whether to focus solely on 2D data or extend to 3D and dynamic imaging. In this section, we classify foundation models for MIS according to their dimensionality, discussing distinct approaches for 2D and higher-dimensional data.

2D images

2D imaging remains one of the most established and widely used formats in medical imaging, typically acquired

through modalities such as X-ray, dermoscopy, fundus photography, ultrasound, and endoscopy. Consequently, foundation models for 2D medical imaging are actively being developed, utilizing various backbone networks, such as U-Net, Transformer, generative adversarial network (GAN) (99), and SAM. For instance, several studies incorporated user-interaction capabilities using established frameworks such as U-Net or Transformer-based models (46,85). In particular, Wu and Xu (85) modified an existing Transformer model to enable user interaction, while UniverSeg (46) introduced a lightweight fusion module into a simple U-Net architecture to incorporate user inputs effectively.

Interestingly, some studies utilized GANs for MIS. For instance, Xu *et al.* (80) proposed Adversarial Privacy-aware Perturbations on Latent Embedding (APPLE), which introduced a small latent feature perturbed without modifying the architecture or parameters of the pretrained foundation model. By perturbing the latent embedding with a GAN, APPLE was designed to improve model fairness in medical segmentation tasks and prevent the leakage of sensitive information to the segmentation decoder.

SAM, known for its exceptional segmentation performance in natural image segmentation, has inspired numerous studies. Studies on SAM-based approaches have explored in various directions, including maintaining the original architecture, modifying the image encoder, refining prompt encoding, adjusting the mask decoder, and making overall structural changes (22,23,93,96). Specifically, one straightforward approach is to retain SAM's entire structure and adapt it to the medical domain (22,44,47). These studies preserved SAM's original architecture and enhanced performance by training some or all of SAM's parameters using medical datasets. In contrast, other studies added modules to the image encoder or replaced the backbone to incorporate domain-specific information (23,48,49,53,60,65,73,74,83,94). These studies primarily focused on incorporating medical domain-specific information by adding modules such as Low-Rank Adaptation (LoRA) (100) or CNN branches to the image encoder. Lin *et al.* (95) supplemented local information by adding a CNN branch to the ViT-based encoder for ultrasound image segmentation. Additionally, ongoing studies modified the structure of the decoder (75,79,96). Gao *et al.* (96) enhanced SAM by modifying its mask decoder and introducing two new components: the prompt-relevant IoU module (PRIM) and the prompt-decoupled mask module (PDMM). PRIM is responsible for generating the IoU score and mask embedding, while PDMM combines multi-scale features from the image encoder with the mask embedding through up-sampling. This structural separation helps preserve pretrained weights as much as possible and minimizes performance degradation caused by incorrect prompts. Lastly, some studies altered the overall structure of SAM (92,93,97). Notably, Dr-SAM (93) and SAM-UNet (97) leveraged the strengths of the U-Net architecture in image segmentation, reconfiguring SAM to adopt a U-Net-like structure.

Higher-dimensional images

3D imaging is primarily obtained through modalities such as CT, MRI, and PET, where volumetric images are constructed from multiple 2D cross-sectional slices. Four-dimensional (4D) imaging extends this concept by incorporating a temporal dimension, as seen in fMRI, 4D CT, and 4D ultrasound, which capture dynamic physiological changes over time. Such images provide essential spatial and temporal data, enabling a comprehensive analysis of anatomical structures and their functional dynamics. While higher-dimensional imaging (e.g., 4D imaging) presents unique challenges, this review focuses on 3D volumetric data. Most foundation models are designed for 3D medical imaging, while higher-dimensional data is often addressed by extending these models or developing specialized architectures tailored to temporal dynamics. By reviewing 3D models, we explored how they process volumetric data and adapt to additional dimensions, such as the temporal axis.

Traditional 2D models work by decomposing 3D volumetric images into 2D slices and processing each slice independently. However, this approach fails to capture the spatial relationships between slices, hindering comprehensive 3D medical image analysis. To address this, recent studies have focused on developing models that process 3D volumetric data. These methods utilized techniques including 3D convolution and 3D patch embedding, along with volumetric segmentation models such as UNEt TRansformer (UNETR) (101), Swin-UNETR (102), and 3D U-NET (8).

The most straightforward approach to processing 3D volumetric data is to design models with 3D architecture (24,25,41,45,55,70,72,77,87,91). Fischer *et al.* (77) proposed a prompt-able UNETR (PUNETR), built upon the UNETR architecture. They introduced class-dependent prompt tokens for both binary and multi-class segmentation. Furthermore, they integrated prompt-able SWin blocks

(PSWin) to facilitate token-dependent class prediction within the attention mechanism. This enhancement allows for more accurate mask predictions tailored to each class. Wang *et al.* (24) proposed SAM-Med3D, which employed 3D operations such as 3D convolution and 3D positional encoding. They utilized a model architecture identical to SAM and trained it from scratch. Du *et al.* (72) proposed SegVol, a foundation model for 3D volumetric image segmentation. To process 3D volumetric data, they adapted SAM's image encoder with a 3D ViT and trained it accordingly. They also designed a zoom-out-zoom-in mechanism, which significantly reduced computational costs while maintaining precise segmentation.

A 2D-to-3D adaptation approach incorporates additional elements or techniques into the 2D models to make them suitable for 3D environments. This approach attempts to effectively capture spatial information in 3D space. Several studies inserted adapters within the image encoder of existing 2D models to integrated 3D information (50,59,62,63,76,78,82). Gong *et al.* (50) proposed a spatial adapter for processing volume data by depth units and a temporal module for handling video data frame-by-frame, enabling existing 2D models to efficiently manage 3D information. These modules were integrated behind the attention blocks of a fixed image encoder, allowing the model to learn 3D spatial patterns while preserving and utilizing the information from the original image encoder. Yan *et al.* (76) proposed AFter-SAM, designed SAM for 3D MIS by integrating LoRA (52) based adapters into the Axial Fusion Transformer (103). These adapters are specifically designed to effectively capture both intra-slice details and inter-slice contextual information. Another study by Gu *et al.* (78) incorporated an external 3D branch into the encoder. To segment bones in various MRI images, they extended the existing 2D SAM model by incorporating a 3D attention network, enabling effective processing of depth information in MRI scans. Furthermore, they utilized a learnable gate to fuse 2D and 3D information, allowing the model to leverage data from both single slices and the entire 3D volume.

Modality coverage

Modality-agnostic models

With the growing accessibility of high-quality public medical datasets, the development of foundational models encompassing different modalities and body regions has expanded (22-24,46,72,92,93,96,97). Ma *et al.* (22) proposed MedSAM, a foundation model designed to support diverse medical imaging modalities and enable universal MIS. The model was trained on a large-scale medical image dataset consisting of 1,570,263 image-mask pairs encompassing 10 imaging modalities and over 30 cancer types. Consequently, MedSAM consistently outperformed state-of-the-art (SOTA) foundation models across 86 internal validation tasks and 60 external validation tasks. Cheng *et al.* (23) proposed SAM-Med2D, which adapted SAM for medical 2D images. They froze the image encoder and integrated learnable adapter layers into each transformer block to acquire domain-specific knowledge for medical applications. They constructed a MIS dataset consisting of over 4.6 million images and 19.7 million masks, approximately 19 times larger than MedSAM's dataset, encompassing various modalities, anatomical structures, and organs. SAM-Med2D outperformed SAM and MedSAM in extensive zero-shot assessments, indicating a substantial improvement in generalization performance. Yang *et al.* (97) proposed SAM-UNet, a foundation model that incorporated U-Net into the original SAM. A parallel convolutional branch was included in the image encoder and trained independently from the transformer branch. For training, they used SA-Med2D-26M dataset (104), which is the largest 2D MIS dataset to date.

Modality-specific models

Modality-specific foundation models are designed to target a single modality. While modality-agnostic foundation models offer the advantage of being broadly applicable across various medical imaging modalities without additional adjustments, these models excel by capturing the unique characteristics of each modality more precisely, thereby delivering an optimized performance. Examples of modality-specific foundation models include dermoscopy-specific model (44), endoscopy-specific models (47,56-58,74,75,88,90), X-ray-specific models (67,81), ultrasound-specific models (62,64,68,83,86,94,95), pathology-specific models (53,65), optical coherence tomography (OCT) specific model (60), microscopy-specific models (59,89), MRI-specific models (54,78,84,87,91,98), and CT-specific models (48,49,51,70,77,82). Cox *et al.* (98) proposed a multi-modal 3D foundation model for neuroimaging segmentation. To enable an accurate and efficient analysis of brain tumors and lesions. They employed a two-stage approach utilizing a ViT. Lin *et al.* (95) introduced SAMUS, an ultrasound-specific adaptation of SAM that transfers its strong feature representation

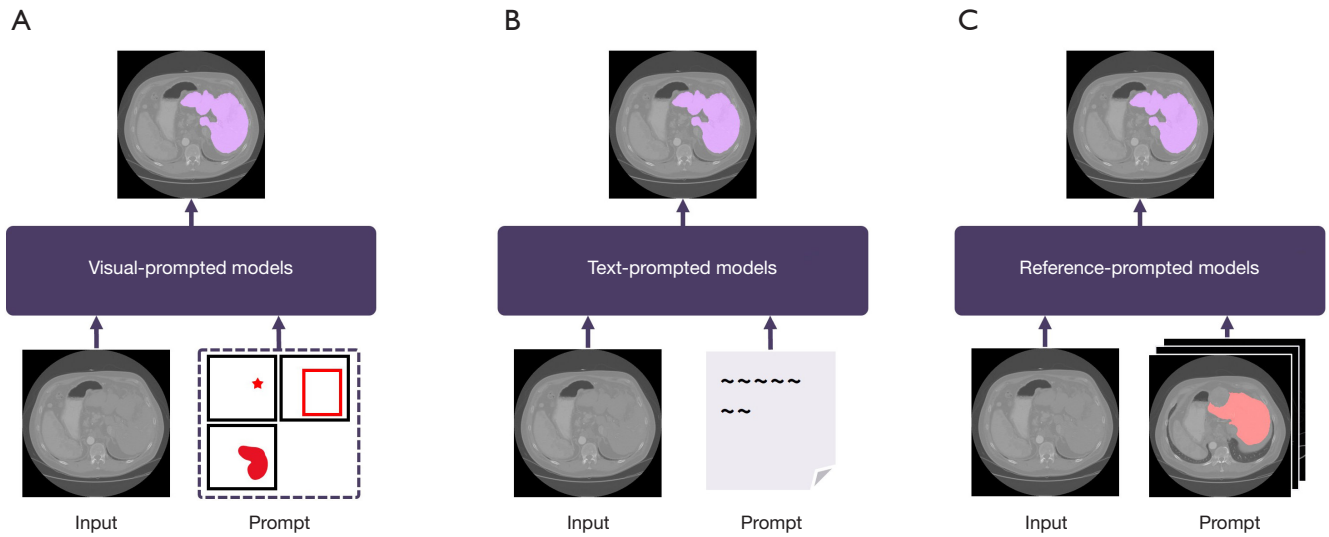


Figure 3 Illustration of models by prompt type. (A) Visual-prompted models using point, bounding box, or mask prompts. (B) Text-prompted models using text, such as target medical terminology or medical reports. (C) Reference-prompted models using reference samples. Medical image was obtained from the publicly available HCC-TACE-Seg dataset (110). HCC, hepatocellular carcinoma; TACE, transarterial chemoembolization.

capabilities to MIS. They developed feature and position adapters for domain adaptation and added a parallel CNN branch to complement local features. A large ultrasound dataset (US30K) was constructed to train and evaluate SAMUS, demonstrating its effectiveness in ultrasound segmentation. Liu *et al.* (65) proposed whole-slide image (WSI)-SAM, a pathology-specific model designed to address the multi-scale nature of WSIs, a key challenge in histopathology. Building upon SAM's prompt-driven, zero-shot framework, WSI-SAM uses multi-resolution patches and minimal additional parameters, including high- and low-resolution tokens in a dual mask decoder, to effectively capture features across scales and enhance segmentation performance. Zhang *et al.* (91) adopted a collaborative approach involving multiple models to perform various tasks, deviating from the typical paradigm of a universal foundation model known as 'One Model for ALL Tasks'. They introduced a novel Mixture of Modality Experts (MoME) framework for segmenting brain lesions across various 3D modalities (e.g., CT, T1, and T2). Inspired by the Mixture of Experts (MoE) (105), MoME comprises modality-specific models designated as experts and employs hierarchical gating to facilitate cooperative processing among these experts. Furthermore, to prevent over-reliance on any single modality expert during training, they incorporated a curriculum learning strategy.

Prompt type

Interactive segmentation has long been a subject of extensive research, focusing on segmenting objects based on user input (15,106-109). Numerous foundation models are designed for interaction with user prompts. As shown in *Figure 3*, these prompts can be categorized into multiple types, which range from visual prompts, such as points, bounding boxes, and scribbles, to semantic prompts such as text. In this section, we focus on the types of prompts utilized by foundation models and their implementation.

Visual prompt

Visual prompts provide direct spatial information about the desired target area using elements such as points, bounding boxes, and scribbles. Due to their intuitiveness and simplicity, these types of prompts have been widely used. Specifically, some studies used point prompts (18,23,24,43,44,50,62,64,67-69,73,78,82,83,86,93,95,96), while others utilized box prompts (18,22,23,43,44,47,49, 62,65,66,71,73,74,78,84,90,93,96,97). Additionally, one study employed scribble-based prompt (74). Scribbles are user-drawn freeform lines that roughly indicate regions of interest, providing a more flexible way to guide segmentation compared to points or boxes. This method enables intuitive annotation of complex shapes, though it

may be less precise for defining exact object boundaries.

Text prompt

While visual prompts provide positional information about the target for segmentation, text prompts offer the model clinical knowledge and contextual information about the target. Several studies (25,58,72,75,81) attempted to leverage the text encoder of large Vision-Language Models such as CLIP (33) to generate text embeddings that capture the prompt's meaning. Specifically, SAT (25) and AdaptiveSAM (58) utilized text prompts to perform segmentation on regions corresponding to specific categories within an image (e.g., pancreas, rib). Furthermore, Du *et al.* (72) introduced a method that combines visual and text prompts. By integrating spatial and semantic information, this approach enables clearer identification of the desired region, reducing ambiguity.

A category-level text prompts (e.g., liver, breast tumor) are convenient but may lack detailed descriptions of the target, which may degrade segmentation accuracy. To address this problem, some studies focused on developing more refined prompts. Yue *et al.* (75) proposed collaborative prompts, which integrate category-level and part-level text for describing target object structures. Using their proposed part-to-whole adaptation fusion module, they effectively integrate detailed information at the part level with category-level embeddings. Killeen *et al.* (81) proposed FluoroSAM, a language-aligned foundation model for X-ray images. FluoroSAM accepts initial descriptors as prompts, including organ names and examples (e.g., L4 vertebra). To increase input diversity, the initial descriptors are augmented using GPT-3.5 Turbo (111) from OpenAI (e.g., lumbar vertebra 4, fourth lumbar vertebra). Finally, the augmented descriptions are embedded using the pretrained MedCLIP (112) model.

Some studies adopted text prompts to generate bounding boxes for target objects specified by the prompts (42,57,88,94), providing detailed textual descriptions of the object's shape, size, and location. Specifically, GSAM+Cutie (88) and Polyp-SAM++ (57) utilized GroundingDINO (113) to derive bounding boxes from text prompts. Meanwhile, Gowda and Clifton (94) also employed GroundingDINO but introduced two major modifications to improve bounding box accuracy: using GPT-4 to generate more descriptive text prompts and replacing the text encoder with MedBERT (114) for better adaptation to the medical domain. In contrast, Koleilat *et al.* (42) introduced a zero-shot medical segmentation approach by fine-tuning BiomedCLIP (115) and applying gScoreCAM (116) to obtain bounding boxes from text prompts. They refined the resulting saliency maps with a conditional random field filter before using these bounding boxes in SAM to produce pseudo-masks.

Reference prompt

Some studies approached this by using reference prompts to segment targets based on other samples (46,61,69,70,85). Butoi *et al.* (46) proposed UniverSeg, which utilized a subset of image-label pairs to identify regions for segmentation in unseen images. These image-label pairs were defined as a support set. To transfer the segmentation task from the support set to the unseen image, which is referred to as the query image, they introduced CrossBlock. This module computed the cross-attention scores between each image-label pair in the support set and the query image, integrating information from the support set to guide segmentation in the query image. While UniverSeg used image-label pairs as a reference, Wu and Xu (85) utilized image-prompt pairs as a support set. They proposed One-Prompt Former, which combined the strengths of one-shot and interactive methods. One-Prompt Former integrated the prompted template features with query features across multiple feature scales. Additionally, through comparative experiments with various prompts such as click, doodle, and bounding box, they evaluated how each prompt can be optimized for specific medical targets, demonstrating its potential to meet diverse clinical requirements.

Training strategy

Supervised pre-training

Supervised learning utilizes large-scale labeled datasets to enhance feature extraction and improve model generalization. It is different from training from scratch in that it leverages a large-scale, labeled dataset to learn general features before fine-tuning, whereas training from scratch initializes random weights and learns solely from the target dataset. By training on extensive annotated datasets, models can learn rich feature representations that contribute to improved performance in MIS. Several studies applied supervised pre-training to improve segmentation performance in medical imaging (45,46,87). Specifically, Butoi *et al.* (46) applied supervised pre-training to UniverSeg to enhance its generalization to downstream tasks. To accomplish this, they constructed a dataset comprising 22,000 scans across 26 medical domains

and 16 imaging modalities, enabling the model to learn diverse anatomical structures and imaging characteristics. Similarly, STU-Net (45) was pretrained on the large-scale TotalSegmentator (117) dataset, which consists of 1,024 CT images with annotations for 104 anatomical structures throughout the entire body.

Self-supervised pre-training

Self-supervised learning (SSL) leverages large-scale unlabeled datasets to enable models to learn meaningful feature representations without requiring labeled data. By leveraging large-scale unlabeled datasets, SSL enables models to learn meaningful feature representations through various pretext tasks. Representative SSL methods include contrastive learning approaches, such as Simple Contrastive Learning Representation (SimCLR) (118) and Momentum Contrast (MoCo) (119), along with masked image modeling techniques, such as MAE. Several studies have employed self-supervised pre-training to enhance MIS performance by leveraging large-scale unlabeled datasets (51,72,77,98). Compared to supervised pre-training, SSL enables the utilization of significantly larger datasets without the need for manual annotations, thereby reducing reliance on expert labeling and improving model generalization across diverse imaging modalities. Specifically, Du *et al.* (72) utilized 90,000 unlabeled CT volumes for pre-training, adopting the SimMIM algorithm (120) with a masked image modeling loss to train its image encoder. Cox *et al.* (98) introduced a two-stage self-supervised pre-training strategy for multimodal neuroimage segmentation. In the first stage, the model learned general anatomical structures from 41,400 unlabeled MRI scans of healthy brains, capturing key features such as shape and size. In the second stage, the model refined its representations by identifying disease-specific patterns, including tumor and lesion geometry and spatial distribution. These approaches demonstrate the potential of SSL in MIS, effectively reducing reliance on annotated datasets while improving generalization across diverse imaging modalities.

Training from scratch

Pre-training on large-scale datasets is a widely adopted strategy in deep learning to enhance feature extraction and improve generalization. However, some studies opted to train models from scratch rather than using pretrained weights (24,41,55,74,89). For instance, Wang *et al.* (24) adopted a fully 3D architecture and trained their model from scratch to address the limitations of other SAM-based approaches. Previous methods, such as 3D adapter techniques, retained a 2D structure while incorporating 3D information in a limited capacity, which restricted their ability to fully leverage. Moreover, SAM's pretrained weights, optimized for 2D natural images, may introduce biases that limit generalization to medical imaging. By training from scratch, they aimed to fully utilize 3D spatial information. Additionally, this approach eliminates biases inherent in 2D pretrained weights, leading to better generalization in medical imaging tasks. Similarly, Gao *et al.* (41) trained MedFormer from scratch, focusing on mitigating the domain gap with natural images and ensuring robust performance under limited medical datasets.

Fine-tuning

Once a model has been pretrained using a supervised or self-supervised approach, Fine-tuning is often applied to adapt it to a specific domain or task. The most straightforward approach to applying foundation models trained on natural images to medical imaging is to fine-tune all parameters of the model for medical imaging. Some studies demonstrated strong performance in the medical domain through full fine-tuning of medical datasets (22,44,45,51,77,87,91,98). However, foundation models contain an enormous number of parameters, making full fine-tuning both time- and resource-intensive. Additionally, updating all parameters poses the risk of forgetting previously learned information.

To overcome these challenges, researchers have explored parameter-efficient fine-tuning (PEFT) (121) to adapt foundation models to specific domains more efficiently. Some studies froze the majority of parameters and only fine-tuned a subset, optimizing the models for medical imaging (47,58,64,84). Recently, the LoRA technique has gained popularity as a PEFT method, with studies leveraging it to build models tailored to the medical domain (48,60,73,76). Wang *et al.* (60) effectively integrated trainable LoRA layers into each block of a fixed image encoder, enhancing OCT angiography images with minimal additional parameters.

Adapter, another approach within the PEFT methodology, is a lightweight module injected into pretrained foundation models. This allows only the adapter modules to be trained for a specific domain, preserving the pretrained model's knowledge. This approach maintains the generalized information learned during pre-training without updating all model parameters. Some studies optimized SAM for the medical domain by inserting and training lightweight adapters within the image encoder or mask decoder (23,54,65,75,76,78,94,95,97). Other studies incorporated

adapters designed to capture 3D spatial or temporal information into 2D models, enabling the processing of 3D volumetric data, such as CT and MRI scans, as well as temporal data such as videos (43,50,62,63).

Additionally, efforts to adapt models for medical applications included the use of trainable lightweight networks. Huo *et al.* (93) proposed a trainable U-shaped residual network and a medical output token to capture unique features at different levels of medical images and improve mask granularity. Other studies developed a trainable CNN branch that operates in parallel with the ViT encoder in the image encoder. This design efficiently captures additional local information relevant to the medical domain (94,95,97). Inspired by the Ladder-Side Tuning (LST) network for Transformers (122), Chai *et al.* (49) integrated a complementary CNN alongside the standard SAM network. This approach allows the flexible integration of an auxiliary network while avoiding backpropagation through the entire large model.

Prompt engineering

Numerous studies refined or enhanced prompts to improve the adaptation of foundation models to MIS (60,67,72,75,81). Wang *et al.* (60) refined SAM's point prompts by automatically sampling paired foreground and background points for each connected vascular component, thereby improving segmentation accuracy. Sathish *et al.* (67) proposed the SAM prompt optimization technique (SAMPOT), an automated prompt tuning method. SAMPOT introduces an oracle that assesses the quality of segmented masks and iteratively updates the prompts to achieve optimal performance.

Prompt-based approaches require user intervention, which is a major limitation in large-scale datasets. This dependence on manual prompts complicates real-time processing and large-scale applications. To address this issue, various approaches developed to enable fully automated prompt generation, allowing models to generate and learn prompts autonomously without the need for user involvement (42,46,52,53,56,57,61,66,69-71,78, 84,84,88,92,94,95).

Some studies utilized localization models for automatic visual prompt generation (57,66,71,84,88,94). Pandey *et al.* (66) applied YOLOv8 (123) to a foundation model to automatically generate bounding box prompts for regions of interest. On the other hand, other studies have proposed additional network training or learnable prompt embeddings. Xu *et al.* (92) proposed the self-patch prompt generator, which contains a patch generator and dense prompt encoder to automatically produce a set of high-quality patch prompts. Yue *et al.* (56) proposed a lightweight prototype-based class prompt encoder, which directly generates prompt embeddings from class prototypes. Furthermore, these prototypes were treated as learnable embeddings, using contrastive learning to clearly distinguish each class prototype. Chen *et al.* (79) proposed a multi-scale self-prompt generation (SPGen) module, which autonomously generates high-quality mask hints and effectively guides the decoder, instead of providing prompts for each region of interest around the nuclei.

Some studies proposed an approach that automatically generates prompts from unlabeled images by referencing a minimally manually prompted support set (46,61,69,70,85). Lei *et al.* (70) proposed the Localize Anything Model for 3D Medical Images (MedLSAM), which can localize any target anatomical structure within the body. To accomplish this, MedLSAM employed two self-supervision tasks: unified anatomical mapping (UAM) and multi-scale similarity (MSS). Using these mapping modules, MedLSAM effectively localized anatomical structures with only a few template scans.

Experiments

In the previous section, we provided a comprehensive overview of studies focused on developing foundation models for the medical domain. However, to effectively apply these models in clinical applications, it is crucial that they demonstrate consistent performance on new datasets. In this section, we outline the key procedures for evaluating model performance, including dataset selection, model selection, prompt generation, data preprocessing, and evaluation metrics.

Experimental setting

The models used in the experiments were selected according to the following criteria. First, we prioritized foundation models that emphasize generalization across modalities. Second, models lacking publicly available weights or adequate reproducibility documentation (e.g., missing essential components such as a configuration file or detailed explanations for text prompt augmentation) were excluded. As illustrated in *Figure 4*, we systematically filtered the initial pool of 63 models based on the defined criteria, resulting in six modality-agnostic foundation

models that met the reproducibility requirements and were used in our experiments. For visual prompts, this includes MedSAM (22), SAM-Med2D (23), and SAM-Med3D (24). For text prompts, this includes SAT-Pro (25), the largest model in the SAT series with 447 million parameters. For hybrid prompt, which combines visual and text input, this includes SegVol (72). For reference prompts, this includes UniverSeg (46). In addition to these six foundation models, the foundation model SAM (15), designed for general image segmentation, was selected as the baseline model. The experiments were conducted at two resolutions: 1,024×1,024 and 256×256, with the latter commonly used in the medical domain. To differentiate between the two resolutions, the SAM model trained at a resolution of 1,024×1,024 is referred to as SAM-1024, while the one trained at 256×256 is referred to as SAM-256. Ultimately, we selected five 2D models, including SAM-256 and SAM-1024, and three 3D models. Detailed information about the selected models is summarized in *Table 3*.

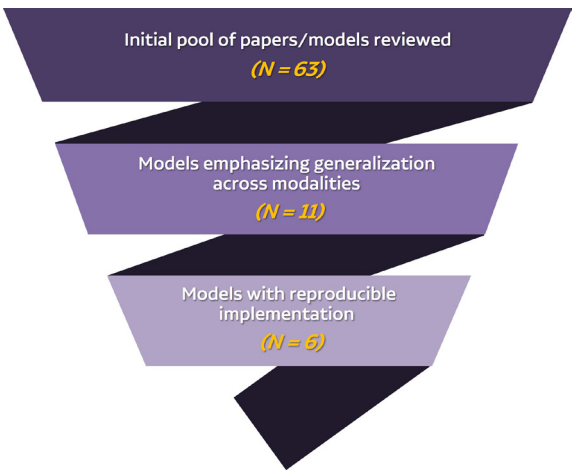


Figure 4 The funnel diagram illustrates the selection process for modality-agnostic models used in comparative experiments. An initial pool of 63 models was narrowed to 11 by prioritizing cross-modality generalization and further reduced to a final 6 by excluding those lacking publicly available weights or sufficient reproducibility guidelines.

After model selection, we conducted a pre-processing phase tailored to the requirements of each model. The most crucial step in this process was the generation of prompts. As SAM-Med3D uses only point prompts, other visual prompt-based models were configured to use point prompts to ensure a valid comparison. Point prompts were generated using the One-point Prompt Generation algorithm illustrated in *Figure 5* and applied to SAM-256 and SAM-1024. Meanwhile, SegVol and SAT-Pro, which use text prompts, employed medical terminology related to the segmentation targets as their text prompts. The class names for the text prompts used in the experiments are listed in Table S1. Lastly, for UniverSeg, two samples were randomly selected and used as the reference prompts.

Dataset

Data curation

We curated a dataset from publicly available sources, primarily sourced from The Cancer Imaging Archive (TCIA) (124), Kaggle (125), Zendo (126), Institute of

Table 3 Summary of the key characteristics of models evaluated for zero-shot segmentation across various dimensions, image domains, dataset sizes, base models, prompt types, and inference input sizes

Dimension	Model	Image domain	Dataset scale	Baseline	Prompt type	Inference input
2D	SAM-256	Natural	1.1B Masks[†]	SAM	Bbox, point	256×256
	SAM-1024	Natural	1.1B Masks[†]	SAM	Bbox, point	1,024×1,024
	MedSAM	Medical	1.1M Masks	SAM	Bbox, point	1,024×1,024
	SAM-Med2D	Medical	19.7M Masks	SAM	Bbox, point	256×256
	UniverSeg	Medical	22K Scans	UNet	Reference	128×128
3D	SAM-Med3D	Medical	131K Masks	SAM	Point	128×128×128
	SegVol	Medical	150K Masks	SAM	Bbox, point, text	*
	SAT-Pro	Medical	302K Masks	SAM	Text	512×512×512

[†], datasets are from the natural domain, not medical datasets. *, SegVol supports all inference input types. 2D, two-dimensional; 3D, three-dimensional; Bbox, bounding box; SAM, segment anything model.

Algorithm 1 One-point Prompt Generation

Input: Binary mask $M \in \{0,1\}^{H \times W}$, Number of points to sample: 1
Output: Sampled point $(x, y) \in R^{1\times 2}$, Label $\ell \in \{0,1\}$ (foreground or background)
Initialize coordinates from the mask *M*:
$fg\_coords \leftarrow \{(x,y) \mid M[x,y]=1\}$ ▷ Foreground coordinates
$bg\_coords \leftarrow \{(x,y) \mid M[x,y]=0\}$ ▷ Background coordinates
Determine size of foreground and background regions:
$fg\_size \leftarrow |fg\_coords|$
$bg\_size \leftarrow |bg\_coords|$
Sample one point:
if $fg\_size > 0$ **then**
  Randomly sample $fg\_coods \in fg\_coords$
  Assignl $\ell \leftarrow 1$ ▷ Foreground label
else
  Randomly sample $bg\_size \leftarrow |bg\_coords|$
  Assignl $\ell \leftarrow 0$ ▷ D Background label
end if
return sampled point (x, y) and label ℓ

Figure 5 One-point prompt generation algorithm. This algorithm generates a single-point prompt from a binary mask M by randomly sampling a foreground $(\ell = 1)$ or background $(\ell = 0)$ point. Foreground and background coordinates are derived from the mask, enabling prompt-based segmentation tasks.

Electrical and Electronics Engineers (IEEE) DataPort (127), and Grand Challenge (128). To evaluate performance on unseen datasets, we excluded datasets that were used for training the 6 models selected for comparative experiments, while maintaining a diverse representation of modalities, anatomical structures, and regions. As a result, we selected 31 datasets comprising 50,966 2D images and 1,603 3D volume scans. *Table 4* summarizes the collected datasets, and *Figure 6* illustrates their composition.

As shown in *Figure 6B*, the curated dataset comprises 10 imaging modalities, such as CT, MRI, X-ray, dermoscopy, endoscopy, and histopathology, with CT images accounting for the largest proportion. Additionally, *Figure 6C* presents the dataset's distribution across four major regions and 16 anatomies, including lesions such as skin cancer and breast tumors. This comprehensive dataset serves a crucial foundation for evaluating the performance and robustness of the models. All datasets used in this study are publicly available and can be accessed through the links provided in Table S2.

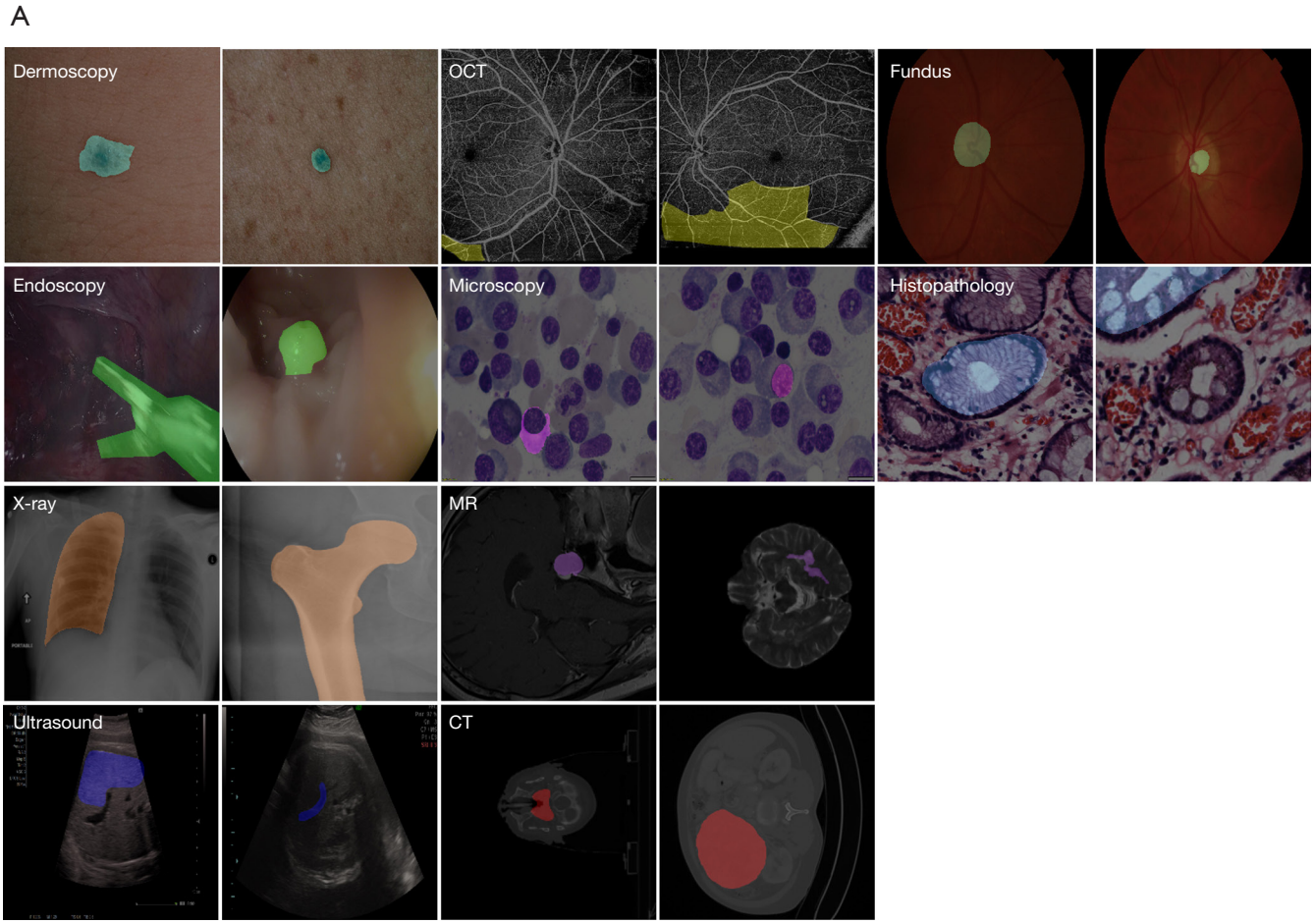
Data pre-processing

The datasets we compiled have different formats, modalities, spatial resolutions, labeling methods, and dimensions. In particular, 3D volumetric data requires transformation to ensure compatibility with 2D models. To standardize the datasets, we conducted pre-processing. First, we standardized the format of all datasets. For 3D volumetric data, such as CT and magnetic resonance (MR), various formats such as Digital Imaging and Communication in Medicine (DICOM), MetaImage Header Archive (MHA), and Nearly Raw Raster Data (NRRD) were converted to the Neuroimaging Informatics Technology Initiative (NIfTI) format. Additionally, all 2D data were converted to the PNG format. Then we performed intensity normalization to improve dataset quality and consistency. Based on Hounsfield unit (HU) values, the 3D volumetric datasets were clipped to minimum and maximum values and then normalized to a range of 0–255. Similarly, the pixel values of the 2D datasets were also normalized to the range of 0 to 255. Subsequently, to compare the performance of 2D models, 3D volume data was sliced along the axial axis. Each mask in the dataset was separated to ensure that only one object was included per mask. For prompt-based models such as SAM and MedSAM, if multiple instances were present, each instance was segmented into a separate mask. Finally, images were resized using bilinear interpolation to match the input resolution required by each model, while

Table 4 Datasets to evaluate zero-shot performance

Dataset	Dimension	Modality	Segmentation targets	# of images/scans
UWater Skin Cancer (129)	2D	Dermoscopy	Skin cancer	180
ETIS-LaribPolypDB (130)	2D	Endoscopy	Polyp	196
AutoLaparo (131)	2D	Endoscopy	Uterus, surgical instruments	1,800
DRAC2022 (132)	2D	OCT	Diabetic retinopathy lesions	106
SegPC2021 (133)	2D	Microscopy	Cytoplasm, nucleus	497
MitoEM (134)	2D	Microscopy	Mitochondria instance	230
PAPILA (135)	2D	Fundus	Optic disc and cup	488
RAVIR (136)	2D	Fundus	Vein, artery	23
PathologyImagesForGlandSeg (137)	2D	Histopathology	Gland	20,000
GlaS@MICCAI2015 (138)	2D	Histopathology	Adenocarcinomas	165
MoNuSAC2020 (139)	2D	Histopathology	Cell	100
WSSS4LUAD (140)	2D	Histopathology	Tissue	120
COVID-19 Radiography (141,142)	2D	X-ray	Lung, COVID-19 infection, lung opacity, viral pneumonia	21,165
ARCADE (143)	2D	X-ray	Coronary artery disease	2997
RBIS-DDSM (144)	2D	X-ray	Breast cancer	689
Xray_hip (145)	2D	X-ray	Femur, ilium	140
QAMEBI (146,147)	2D	US	Benign & malignant breast lesion	232
BUSC (148)	2D	US	Benign & malignant breast lesion	250
USFetalHead (149)	2D	US	Stomach, liver, vein	1,588
TDSC-ABUS2023 (150)	3D	US	Breast tumor	100
RESECT-SEG (151)	3D	US	Brain tumor, resection cavity	69
ATLAS2023 (152)	3D	MR	Liver, tumor	60
CrossMoDA2022 (153)	3D	MR	Brain tumor, cochlea	210
PASeg (154)	3D	MR	Pituitary adenoma	55
Shifts2022 (155,156)	3D	MR	White matter MS lesion	371
ASOCA (157)	3D	CT	Coronary artery	40
Adrenal-ACC-Ki67-Seg (39)	3D	CT	Adrenal mass	53
HCC-TACE-Seg (110)	3D	CT	Liver	104
ImageTBAD (158)	3D	CT	Aorta	100
InnerEarSeg (159)	3D	CT	Inner ear	341
SegRap2023_Task2 (160)	3D	CT	GTVp, GTVnd	100

2D, two-dimensional; 3D, three-dimensional; CT, computed tomography; GTVnd, lymph node gross tumor volume; GTVp, primary gross tumor volume; MR, magnetic resonance; MS, multiple sclerosis; OCT, optical coherence tomography; US, ultrasound.



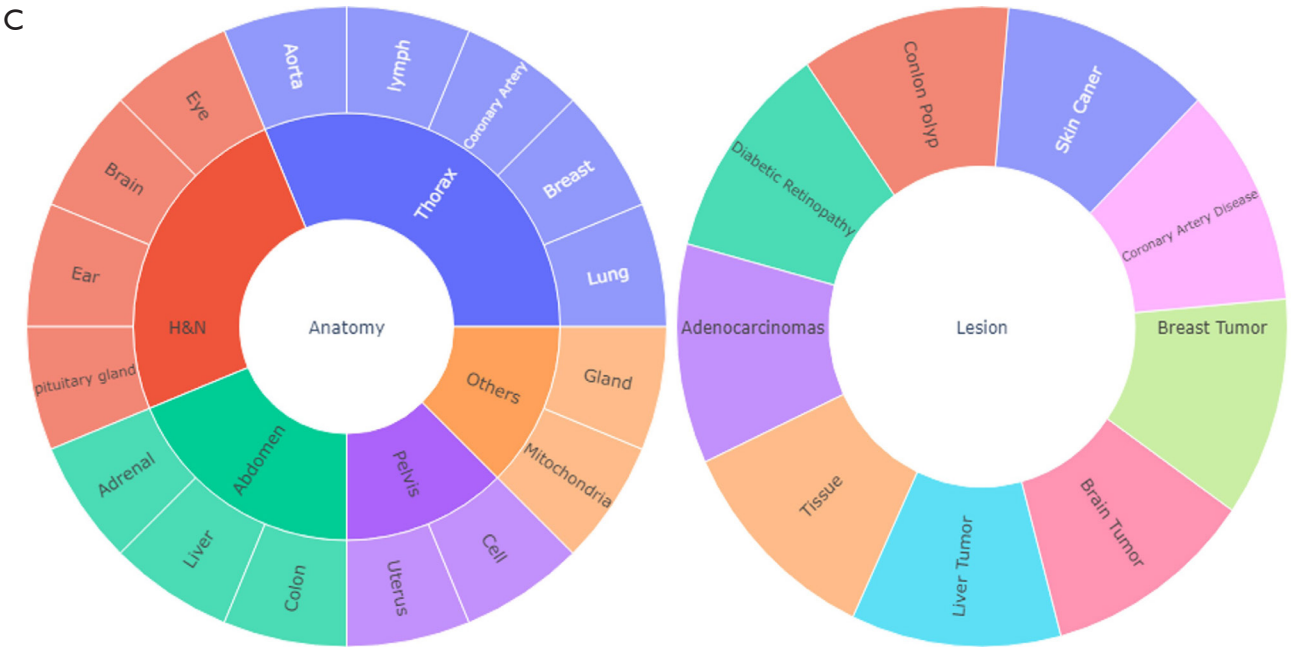


Figure 6 Overview of diverse datasets for MIS in this study. (A) Representative examples of imaging modalities and their zero-shot segmentation targets. All images were sourced from publicly available datasets: dermoscopy—UWater Skin Cancer (129); endoscopy—AutoLaparo (131); OCT—DRAC2022 (132); microscopy—SegPC2021 (133); fundus—PAPILA (135); histopathology—PathologyImagesForGlandSeg (137); X-ray—COVID-19 Radiography (141,142); ultrasound—USFetalHead (149); MR—PASeg (154) and Shifts2022 (155,156); and CT—Adrenal-ACC-Ki67-Seg (39) and SegRap2023_Task2 (160). Colored overlays denote predicted structures: cyan for skin cancer lesions; green for surgical instruments and uterus; yellow for diabetic retinopathy lesions; magenta for nuclei; mint green for optic disc and cup; light blue for cells; orange for lung fields; blue for liver and fetal head vein; violet for pituitary adenoma and vestibular schwannoma; and red for adrenal gland and primary gross tumor volume GTVp. (B) The number of images in each modality. (C) Anatomical distribution across different regions. CT, computed tomography; H&N, head & neck; GTVp, primary gross tumor volume; MIS, medical image segmentation; MR, magnetic resonance; OCT, optical coherence tomography.

masks were resized using nearest-neighbor interpolation. Masks that covered less than 0.456 of the total image area were excluded from the experiments.

Evaluation metrics

To quantitatively evaluate the segmentation performance, we utilized two widely used evaluation metrics in the field of image segmentation: Dice similarity coefficient (DSC) (161) and normalized surface distance (NSD) (162). DSC is a region-based evaluation metric that measures the overlap between the expert-annotated mask G and the model's predicted mask P, which is defined as:

$$DSC(G,P)=\frac{2|G\cap P|}{|G|+|P|} \quad [1]$$

The DSC value ranged from 0 to 1, with values closer to 1 indicating a higher degree of agreement between the two masks.

NSD is a boundary-based evaluation metric that quantifies how well the boundary regions of the expert-annotated mask G align with those of the model's predicted mask P, which is defined as:

$$NSD(G,P)=\frac{|\partial P\cap B_{\partial G}|+|\partial G\cap B_{\partial P}|}{|\partial P|+|\partial G|} \quad [2]$$

Where $B_{\partial P}$ and $B_{\partial G}$ represent the boundary regions within a given tolerance τ, defined as follows:

$$B_{\partial P}=\left\{x\in\mathbb{R}^d \mid \exists\hat{x}\in\partial P, \left\|x-\hat{x}\right\|\le\tau\right\} \quad [3]$$

$$B_{\partial G}=\left\{x\in\mathbb{R}^d \mid \exists\hat{x}\in\partial G, \left\|x-\hat{x}\right\|\le\tau\right\} \quad [4]$$

In this study, we set the tolerance τ as 2, where d denotes the dimensionality of the data space.

Results

Overall performance

Figure 7 shows the DSC and NSD scores across 31 tasks. We divided the entire dataset into four main groups to analyze performance in more detail and to prevent any one group with excessive data from overshadowing the performance of others in the overall evaluation. The groups were categorized into Red-Green-Blue (RGB) images, grayscale images, CT, and MR. The RGB images consist of dermoscopy, endoscopy, fundus, microscopy, and histopathology, while the grayscale images, including X-ray and ultrasound, are further divided into 2D and 3D subsets. *Table 5* presents the DSC results for these five groups (RGB, grayscale 2D, grayscale 3D, CT, and MR). SAM-1024 achieved the highest performance in the RGB image group with a score of 0.485, while medical foundation models outperformed it in the grayscale image, CT, and MR groups. Specifically, in the grayscale 2D and 3D groups, MedSAM and SAM-Med2D demonstrated performances of 0.637 and 0.399 for 2D and 3D, respectively. In the CT group, SAM-Med2D achieved the best performance at 0.491, and in the MR group, it recorded the highest score of 0.562. *Table 6* shows the NSD results for the five groups, where SAM-Med2D demonstrated the highest performance across all groups except for grayscale 2D images. These results clearly illustrate how model performance varies depending on the image type of each dataset.

Performance evaluation on different modalities

To gain deeper insights, we analyzed the performance of eight foundation models by modality. *Figure 8A* presents the DSC performance of five 2D models across eight distinct 2D modalities. Overall, SAM-256 exhibited the lowest performance in most modalities, including dermoscopy, fundus, histopathology, and X-ray. In contrast, SAM-1024 demonstrated above-average performance across most modalities, except for fundus and ultrasound. MedSAM showed strong adaptability to the medical domain, recording the highest performance in the dermoscopy, fundus, and X-ray modalities, with scores of 0.911, 0.556, and 0.681, respectively. Likewise, SAM-Med2D exhibited consistent and above-average performance across all 2D modalities, achieving a notable 0.727 in the ultrasound modality. Meanwhile, UniverSeg, which uses reference samples as prompts, generally showed weaker performance, particularly underperforming in the endoscopy (mean: 0.455) and microscopy (mean: 0.491) modalities, with scores of only 7.2% and 8.7%, respectively. *Figure 8B* presents the DSC performance of eight models across three different 3D modalities. SAM-Med2D achieved the highest performance in the ultrasound, CT, and MR modalities, with scores of 0.399, 0.491, and 0.562, respectively, demonstrating particularly strong results in MR and CT. Following SAM-Med2D, SAM-Med3D, MedSAM, and SAM-1024 recorded the second-highest scores, with US: 0.310, MR: 0.340, and CT: 0.406. Interestingly, the 3D models, including SAM-Med3D, SegVol, and SAT-Pro, performed worse compared to the 2D models that processed the data slice-by-slice. SAT-Pro, which exclusively uses text prompts, recorded the lowest performance across all modalities, with scores of 0.054 for ultrasound, 0.160 for CT, and 0.033 for MR, respectively.

Performance evaluation on different anatomical structures and lesions

We analyzed model performance by categorizing it across various anatomical structures, types of lesions, and surgical instruments. This analysis helped us identify the strengths of each model in specific anatomical structures and recognize their weaknesses in certain shapes or colors. *Figure 9* displays the DSC performance for 18 anatomical structures, 9 types of lesions, and surgical instruments. For the liver, and adrenal, the 3D models SAM-Med3D and SAT-Pro recorded the highest performance. SAM-1024 showed strong performance across various anatomical structures, achieving scores of 0.664 for the uterus (X-ray), 0.635 for the cytoplasm, 0.922 for the nucleus, 0.843 for the mitochondria, 0.470 for the cell, 0.757 for the inner ear (CT), and 0.837 for the coronary artery (CT). MedSAM achieved top performance in the eye and lung, with scores of 0.769 and 0.836, respectively. UniverSeg outperformed the others in artery, ilium, femur, and portal vein, achieving 0.681 for portal vein (mean: 0.186). For liver and adrenal, SAT-Pro and SAM-Med3D achieved the highest scores, at 0.806 and 0.822, respectively, highlighting the strength of the 3D models. Notably, for the liver, all 3D models consistently outperformed the 2D models, whose mean performance was approximately 40%, demonstrating

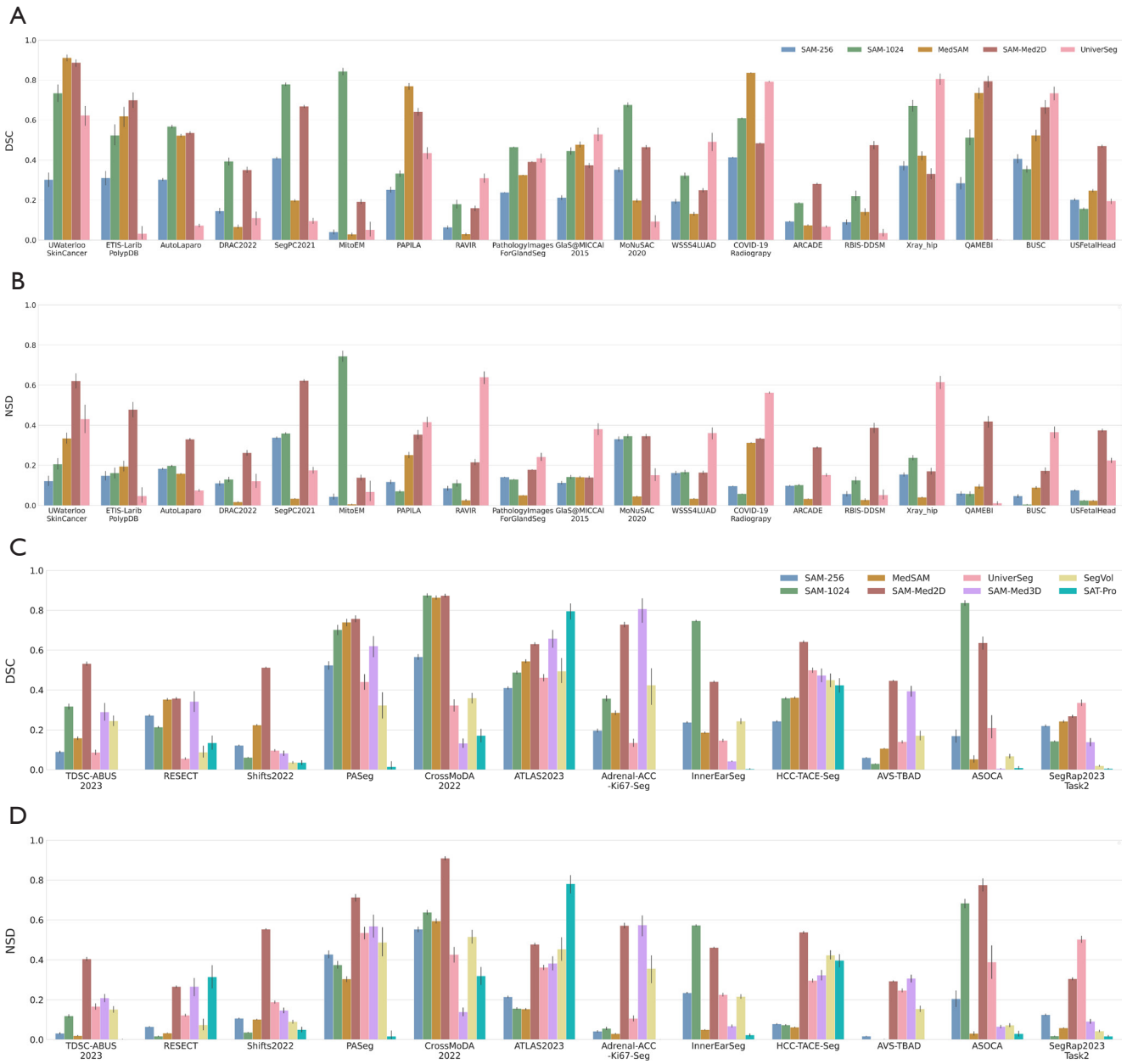


Figure 7 Quantitative performance across different datasets: (A) DSC score on 2D images. (B) NSD score on 2D images. (C) DSC score on 3D images. (D) NSD score on 3D images. For 3D volumetric data, 2D models (SAM-256, SAM-1024, MedSAM, SAM-Med2D, UniverSeg) perform segmentation by slicing the volume into 2D slices and processing each slice individually. 2D, two-dimensional; 3D, three-dimensional; DSC, Dice similarity coefficient; NSD, normalized surface distance; SAM, segment anything model.

superior and consistent results.

In the section highlighted by the yellow box in *Figure 9*, the medical foundation models demonstrated strong performance in most lesion-related tasks, except for diabetic retinopathy. Specifically, MedSAM recorded a score of 0.911 in skin cancer, whereas SAM-Med2D achieved the highest performance in colon polyp, coronary artery disease, breast tumor, and brain tumor, with scores of 0.700, 0.281, 0.416, and 0.510, respectively. Additionally, UniverSeg showed the best performance in adenocarcinomas and tissue regions. In

Table 5 DSC scores of foundation models across five groups: RGB, grayscale 2D, grayscale 3D, CT, and MR

Model	DSC				
	2D		3D		
	RGB	Grayscale	Grayscale	CT	MR
SAM-256	0.249	0.334	0.229	0.186	0.219
SAM-1024	0.485*	0.492	0.238	0.406	0.213
MedSAM	0.328	0.637*	0.307	0.211	0.340
SAM-Med2D	0.411	0.447	0.399*	0.491*	0.562*
UniverSeg	0.145	0.511	0.063	0.233	0.213
SAM-Med3D	–	–	0.310	0.285	0.206
SegVol	–	–	0.182	0.250	0.220
SAT-Pro	–	–	0.054	0.125	0.213

*, highest performance for each group. 3D models (SAM-Med3D, SegVol, and SAT-Pro) were not evaluated on 2D datasets. 2D, two-dimensional; 3D, three-dimensional; CT, computed tomography; DSC, Dice similarity coefficient; MR, magnetic resonance; RGB, Red-Green-Blue; SAM, segment anything model.

Table 6 NSD scores of foundation models across five groups: RGB, grayscale 2D, grayscale 3D, CT, and MR

Model	NSD				
	2D		3D		
	RGB	Grayscale	Grayscale	CT	MR
SAM-256	0.155	0.095	0.056	0.123	0.160
SAM-1024	0.150	0.065	0.041	0.245	0.100
MedSAM	0.058	0.231	0.029	0.038	0.141
SAM-Med2D	0.210*	0.328	0.298*	0.428*	0.557*
UniverSeg	0.143	0.400*	0.133	0.264	0.261
SAM-Med3D	–	–	0.231	0.214	0.196
SegVol	–	–	0.120	0.233	0.295
SAT-Pro	–	–	0.125	0.125	0.261

*, highest performance for each group. 3D models (SAM-Med3D, SegVol, and SAT-Pro) were not evaluated on 2D datasets. 2D, two-dimensional; 3D, three-dimensional; CT, computed tomography; MR, magnetic resonance; NSD, normalized surface distance; RGB, red-green-blue; SAM, segment anything model.

contrast, the 3D model SAT-Pro excelled in the liver and recorded the highest score of 0.661 for the liver tumor.

Qualitative results

We conducted a qualitative analysis of the segmentation results for the eight selected foundation models. *Figure 10* illustrates the visualized results, showing the predicted masks from each model overlaid on the original images for a randomly chosen sample from each modality. *Figure 10A* presents the masks predicted by the 2D models for the 2D datasets. SAM-Med2D generally produced masks that closely resembled the ground truth (GT) across all modalities. However, for the ultrasound images, it failed to accurately capture the target region, highlighting ongoing challenges in segmenting low-contrast and artifact-prone data. SAM-1024 demonstrated strong performance in predicting masks closely aligned with the GT for most

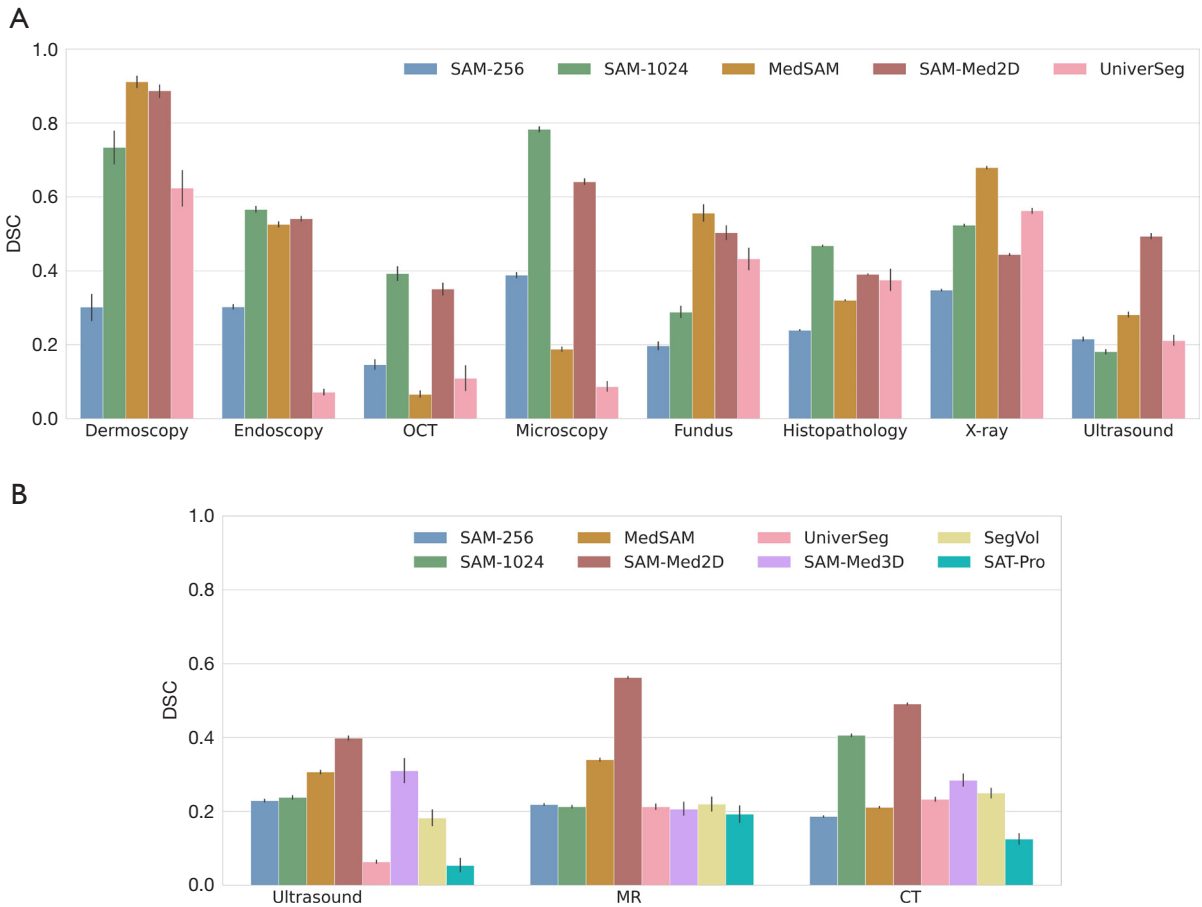


Figure 8 Mean DSC scores across different imaging modalities. (A) Performance on 2D modalities such as dermoscopy and endoscopy. (B) Performance on 3D modalities including ultrasound, MR, and CT. CT, computed tomography; DSC, Dice similarity coefficient; MR, magnetic resonance; OCT, optical coherence tomography; SAM, segment anything model.

RGB image types, including dermoscopy, microscopy, fundus, and histopathology. However, its performance was less effective for endoscopy and fundus. Specifically, in endoscopy, the model failed to detect the polyp fully due to its small size and indistinct contours. Conversely, SAM-256 at a resolution of 256×256 produced inaccurate masks across most modalities. UniverSeg also failed to detect structures completely in endoscopy, OCT, and microscopy.

Figure 10B shows the visualization results for the 3D datasets. For ultrasound data, SAM-1024 successfully generated a mask for the brain tumor region that was highly similar to GT. In contrast, SAM-Med2D, UniverSeg, SegVol, and SAT-Pro did not produce accurate masks. For MR data, most models, except SegVol, created masks closely matching the GT due to the well-defined contours of the regions. Similarly, for CT data, most models accurately predicted masks that matched the GT, thanks to the clear boundaries of the liver, although SAM-256 and MedSAM struggled to delineate the liver region correctly.

An interesting observation is that all 2D models misclassified the liver tumor region as liver, while the 3D models, including SAM-Med3D, SegVol, and SAT-Pro, correctly identified and excluded the liver tumor area. This suggests that the additional spatial context provided by 3D models improves segmentation accuracy by compensating for information loss in slice-based views. Additional visualizations are provided in Figure S1.

Modality-agnostic models vs. modality-specific models

To investigate whether specialized models outperform general-purpose approaches, we conducted additional

Figure 9 Segmentation results for 18 anatomical structures, 9 lesion types, and various surgical instruments. DSC, Dice similarity coefficient; SAM, segment anything model.

comparison experiments between modality-specific and modality-agnostic models. Among modality-specific foundation models, CT and histopathology were the only modalities with publicly available code and pretrained weights, ensuring reproducibility. For each modality, we selected WSI-SAM (65) as a histopathology-specific model, and MA-SAM (63) and STU-Net (45) as CT-specific models. *Table* 7 shows the DSC-based performance comparison among these models. For CT image segmentation (liver and aorta), STU-Net achieved DSC scores of 0.913 and 0.610, while MA-SAM scored 0.837 and 0.529, respectively—both outperformed the modality-agnostic models. For histopathology segmentation (cell, tissue, and adenocarcinoma), WSI-SAM achieved DSC scores of 0.829, 0.868, and 0.868, respectively, thereby surpassing the modality-agnostic models across all tasks. These results indicate that modality-specific architectures more effectively capture the unique features and patterns of their respective imaging domains, leading to superior segmentation performance. By tailoring network components to the characteristics of CT or histopathology, modality-specific models highlight crucial details that general-purpose solutions might overlook.

However, as this study evaluates only a subset of

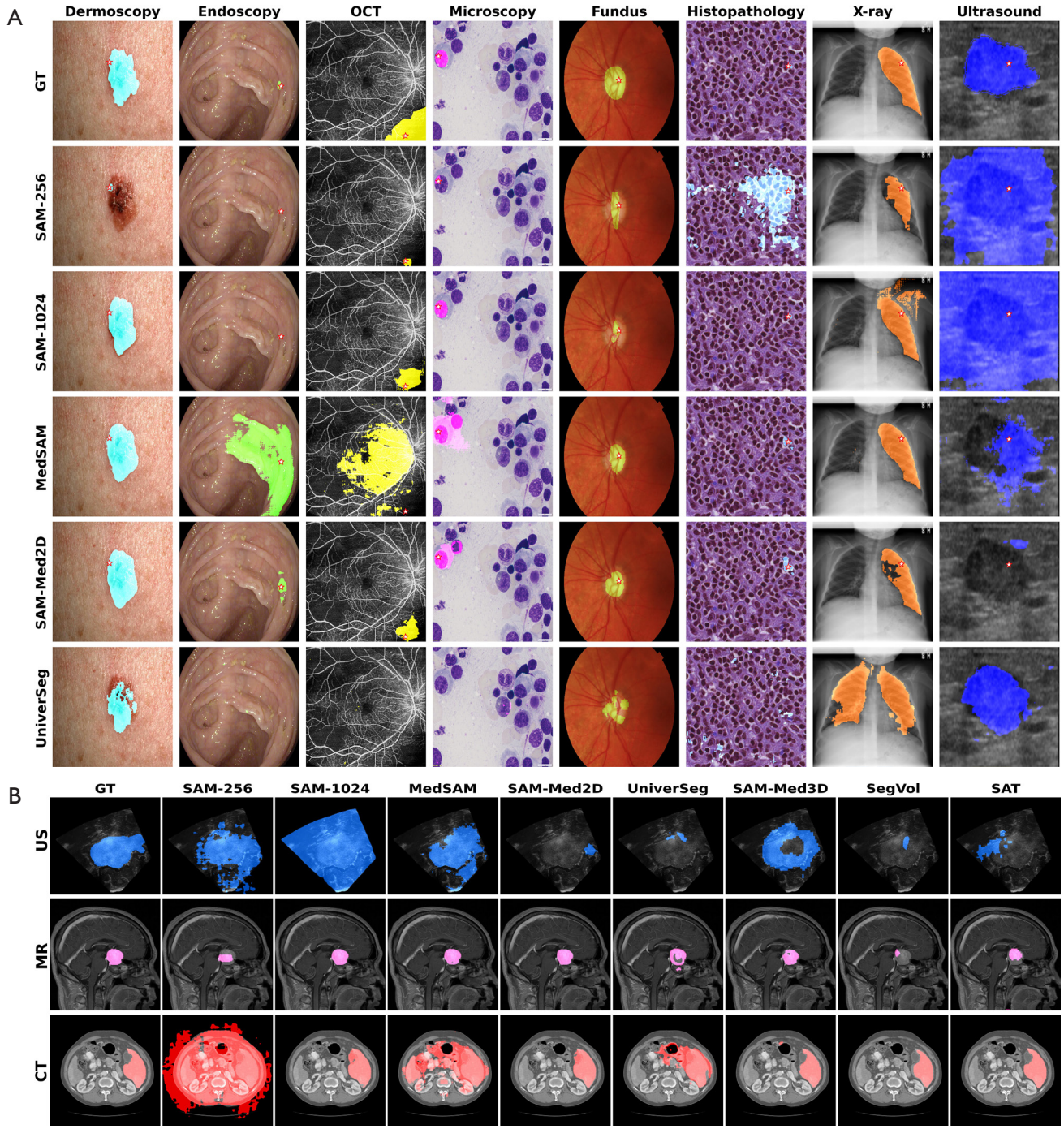


Figure 10 Representative visualization of model predictions across various imaging modalities. (A) 2D imaging modalities. The star indicates the positive point prompt used for interactive segmentation. All medical images were obtained from the publicly available datasets: dermoscopy—UWater Skin Cancer (129); endoscopy—ETIS-LaribPolypDB (130); OCT—DRAC2022 (132); microscopy—SegPC2021 (133); Fundus—PAPILA (135); histopathology—MoNuSAC2020 (139); X-ray—COVID-19 Radiography (141,142); ultrasound—BUSC (148). (B) 3D imaging modalities. All medical images were obtained from the publicly available datasets: ultrasound—RESECT-SEG (151); MR—PASeg (154); CT—HCC-TACE-Seg dataset (110). Cell image dataset specifications: MoNuSAC2020 (40× magnification, H&E stained), SegPC2021 (Jenner-Giemsa stained). CT, computed tomography; H&E, hematoxylin and eosin; GT, ground truth; MR, magnetic resonance; OCT, optical coherence tomography; SAM, segment anything model; US, ultrasound.

Table 7 Performance evaluation of modality-specific models (WSI-SAM, MA-SAM, STU-Net) and modality-agnostic models across various anatomies and lesions. Values represent DSC score

Modality	Model	CT		Histopathology		
		Liver	Aorta	Cell	Tissue	Adenocarcinomas
CT	STU-Net	0.913*	0.610*	–	–	–
	MA-SAM	0.837	0.529	–	–	–
Histopathology	WSI-SAM	–	–	0.829*	0.492*	0.868*
	SAM-256	0.338	0.083	0.241	0.193	0.212
Modality-agnostic	SAM-1024	0.419	0.088	0.47	0.322	0.445
	MedSAM	0.795	0.127	0.321	0.131	0.478
	SAM-Med2D	0.667	0.472	0.393	0.249	0.351
	UniverSeg	0.444	0.166	0.233	0.491	0.529
	SAM-Med3D	0.770	0.405	–	–	–
	SegVol	0.803	0.201	–	–	–
	SAT-Pro	0.770	0.405	–	–	–

*, highest performance for each segmentation target. CT, computed tomography; DSC, Dice similarity coefficient; MA, modality-adaptive; SAM, segment anything model; WSI, whole-slide image.

modality-specific models on CT and histopathology, further research is required to determine whether similar performance trends hold across other medical imaging modalities. Future work should extend this comparison to a broader range of modalities, such as ultrasound, MRI, and endoscopy, to assess the generalizability of modality-specific models.

Discussion

The recent success of foundation models in general domains has sparked significant interest in the medical field, leading to various attempts to develop foundation models tailored specifically for medical applications. Building on this background, we provided a comprehensive overview of research on foundation models for MIS and conducted zero-shot performance evaluations on eight selected models. Through this process, we identified several key findings.

The importance of large-scale medical datasets for foundation models

The scale and diversity of datasets are crucial factors in building foundation models. A dataset that covers a broad range of anatomical structures and imaging modalities can enhance generalization and robustness. As shown in Table S3, most models introduced in this study have collected a variety of public and private datasets and further developed new large-scale datasets for training. For instance, Ma *et al.* (22) developed a dataset comprising around 1.1 million medical image-mask pairs. Ravishankar *et al.* (68) created a dataset with over 0.2 million ultrasound image-mask pairs. Notably, SAM-Med2D, which demonstrated outstanding performance in most cases this study, used 19.7 million masks for training, making it the model trained on the largest dataset among the six foundation models compared.

Our experimental results reinforce the critical role of dataset scale and diversity in model performance, as models trained on larger, more comprehensive datasets demonstrated superior generalization across various imaging modalities. *Figure 8* illustrates this trend, showing that SAM-Med2D exhibited strong performance in both 2D and 3D segmentation tasks, likely benefiting from its extensive dataset coverage. Similarly, MedSAM, trained on 1.1 million samples, performed competitively, particularly in grayscale anatomical structures.

Despite these efforts, the scale of medical datasets still falls short compared to natural image datasets, such as ImageNet (163), JFT-300M (164), LVD-142M (37), and

SA-1B (15). To overcome this limitation, efforts to expand medical datasets through dataset sharing and community building are essential.

2D models vs. 3D models

Although various 3D models have been developed to process 3D volume data such as CT, MR, our performance evaluation revealed that these 3D models fell short of expectations. Specifically, their results were only on par with the baseline natural image foundation models, SAM-256 and SAM-1024, rather than exceeding them. One possible reason for this result is that while the 2D models utilized different 1-point prompts for each slice, the 3D models employed only a single point prompt for the entire scan, potentially limiting the amount of information used. These findings highlight the need for further advancements and improvements in 3D model performance.

Despite their limitations, 3D models demonstrate a clear advantage by incorporating additional spatial context, which improves segmentation accuracy. For example, while 2D models often struggle to differentiate regions such as liver tumors from the liver, 3D models leverage volumetric information to achieve better accuracy.

Our findings suggest that the observed performance gap is not an inherent limitation of 3D models but rather a result of suboptimal prompting strategies and dataset constraints. Future research should explore enhanced multi-slice prompting techniques, hybrid 2D–3D architectures, or adaptive prompting mechanisms to better utilize the spatial information in volumetric data and improve 3D model effectiveness.

Challenges and advances in using prompts for foundation models

Many foundation models utilize prompts, with initial approaches predominantly employing visual prompts (e.g., points, bounding boxes, masks). More recently, research has increasingly explored the integration of text prompts. However, current methods for using text prompts remain relatively simple, and performance evaluations reveal that the text-based model, SAT-Pro, exhibited lower performance across most modalities, except for the liver and liver tumor. A plausible explanation for the model's suboptimal performance is its difficulty in accurately parsing and mapping complex medical terminology to the corresponding anatomical regions. Terms such as "white matter lesion" or "coronary artery" demand specialized domain knowledge, rendering them more challenging to interpret than simpler keywords like "liver" or "lung". This underscores the importance of prompt engineering—particularly in the selection and structuring of terminology—when employing text-based prompts. Moreover, it highlights the need for constructing more anatomically detailed datasets that capture the nuanced variations among different organs and lesions. As these findings indicate, text prompts in medical imaging are still in an early stage of development, suggesting substantial potential for future enhancements in handling domain-specific vocabulary and subtle anatomical distinctions.

The reference prompt also warrants attention. In fact, foundation models utilizing reference prompts have emerged (46,61,69,70,85), aiming to enhance segmentation performance by leveraging a reference sample as guidance. Our performance evaluation showed that UniverSeg demonstrated strong performance for structures with fixed positions in the image, such as the ilium, femur, liver, and portal vein. However, for structures with variable positions, such as the uterus, cytoplasm, and nucleus, the performance was significantly lower. This limitation likely arises from the fundamental nature of reference-based prompting. Unlike point- or bounding-box-based prompts, which provide localized segmentation cues, reference prompts rely on feature similarity between the reference sample and the target image. While this approach is effective for organs and structures that maintain consistent spatial positioning and morphological characteristics, it struggles with structures exhibiting high inter-subject variability in shape, size, or location.

Adaptation

The question of whether SAM can be directly applied to the medical domain has been a subject of ongoing discussion. However, the zero-shot performance results from this study indicate that SAM-1024, despite not being specifically tailored for the medical domain, demonstrated satisfactory performance on 2D medical imaging modalities. Notably, in endoscopy, microscopy, and histopathology, it outperformed foundation models optimized for the medical domain, achieving the highest performance. This may be because these three modalities share characteristics with RGB images commonly used in general domains. Another factor is the lack of training datasets for these modalities. An examination of the dataset distribution reveals that SAM-

Med2D had relatively fewer datasets for these modalities, while MedSAM had significantly limited data for pathology and fundus.

The 3D medical images, which differ significantly from natural images, presented different outcomes. In the CT, US, and MR modalities, foundation models specialized for the medical domain demonstrated superior performance. These results emphasize the importance of adapting models to the medical domain and highlight the need for developing design and training strategies that can more accurately and efficiently process 3D data in the future.

Input resolution & model capacity

Higher-resolution models generally yield better segmentation performance. SAM-1024 consistently outperforms SAM-256 across multiple modalities, likely due to its ability to capture finer spatial details and improve boundary delineation. However, this advantage comes at the cost of increased computational demands. Balancing resolution with efficiency is essential, and optimizing model input size based on anatomical complexity may help maintain segmentation quality while minimizing computational overhead.

Comparison with related studies

Several prior studies have explored advancements within the medical domain. Azad *et al.* (165) provided a comprehensive review of foundation models, categorizing them into textually prompted and visually prompted models across 40 medical datasets. While this study discussed the potential, applicability, and future research directions of foundation models, it did not offer a detailed performance comparison, making it difficult to evaluate their clinical utility. On the other hand, our study directly compares the performance of various models on the same dataset, facilitating a more practical assessment of each model's effectiveness in real-world settings. This comparison provides valuable insights for future researchers when selecting the most suitable model for their specific needs.

Mazurowski *et al.* (166) assessed the real-world generalization performance of SAM in medical contexts by analyzing variations in performance across different prompt types, datasets, and backbone networks. However, their work was limited to using only SAM and datasets such as X-ray, US, MR, and CT. Differently, our study evaluates six foundation models specifically designed for medical applications, broadening the scope to include additional modalities such as dermoscopy, histopathology, ultrasound, and CT. By utilizing 31 unseen datasets spanning 10 modalities, we provide a more comprehensive evaluation of model generalization performance.

Strengths & limitations

This part outlines the limitations and strengths of our research survey on foundation models for MIS, providing insights into the scope of the study and areas for future improvement.

Limitations

(I) This study focuses exclusively on foundation models trained primarily on medical imaging data for segmentation tasks. Models trained on non-medical datasets, such as cross-domain vision-language models, were excluded from our scope, as our primary objective was to analyze models explicitly developed or adapted for MIS. While cross-domain approaches have shown potential for medical applications, systematically evaluating their applicability to segmentation remains an important direction for future research.

(II) The experimental comparison was limited to eight foundation models due to constraints in publicly available model weights and reproducibility. Including a broader set of models in future evaluations could provide deeper insights into their generalization capabilities.

Strengths

(I) The study introduces a distinctive taxonomy based on key criteria such as data dimensions, modality coverage, prompt type, and training strategy to systematically analyze the application of foundation models in the medical domain. This taxonomy may serve as a useful tool for comparing and understanding foundation models within the medical domain.

(II) By conducting zero-shot evaluations on 31 unseen datasets, we provide a comprehensive analysis of model generalizability across various medical imaging modalities and anatomical structures, offering valuable insights into real-world applicability.

Conclusions

This study provides a comprehensive evaluation of

foundation models tailored for MIS. Through zero-shot performance comparisons on 31 unseen datasets across nine imaging modalities, we identified key factors that influence model performance, including dataset scale, input resolution, prompting strategies, and model architecture. Notably, SAM-Med2D and MedSAM demonstrated superior generalization across various 2D modalities, while domain-specific models outperformed general-domain models in 3D tasks.

Our findings highlight the importance of developing large, diverse, and modality-balanced medical datasets. Optimizing prompt designs—particularly for text and reference-based inputs—and tailoring model architectures to the specific characteristics of medical images are also critical for achieving high segmentation performance. Furthermore, our results suggest that models pretrained on general-domain data, such as SAM, can still exhibit competitive performance in medical settings, especially for modalities that resemble natural images (e.g., endoscopy, microscopy, and histopathology). This underscores the potential for cross-domain transfer, while also emphasizing the need for domain adaptation strategies when dealing with complex volumetric data such as CT or MRI.

Future research directions include the development of hybrid 2D–3D models, enhanced prompting techniques (e.g., multi-slice and anatomical-context-aware prompts), and systematic benchmarking of cross-domain vision-language foundation models. These efforts will be essential for advancing the generalizability, clinical reliability, and practicality of foundation models in real-world medical applications.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the Narrative Review reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-2024-2826/rc

Funding: This work was supported by Kyonggi University Research Grant 2024.

Conflicts of Interest: Both authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-2024-2826/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.
2. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, van Ginneken B, Madabhushi A, Prince JL, Rueckert D, Summers RM. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE Inst Electr Electron Eng 2021;109:820-38.
3. Held K, Rota Kops E, Krause BJ, Wells WM 3rd, Kikinis R, Müller-Gärtner HW. Markov random field segmentation of brain MR images. IEEE Trans Med Imaging 1997;16:878-86.
4. LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, Jackel LD. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation 1989;1:541-51.
5. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference. Cham: Springer International Publishing; 2015:234-41.
6. Xiao X, Lian S, Luo Z, Li S. Weighted Res-UNet for High-Quality Retina Vessel Segmentation. 2018 9th International Conference on Information Technology in Medicine and Education; 2018:327-31.
7. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++:

Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. IEEE Trans Med Imaging 2020;39:1856-67.
8. Çiçek Ö, Abdulkadir A, Lienkamp, SS, Brox T, Ronneberger O, editors. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021. Cham: Springer International Publishing; 2021:424-32.
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is All you Need. Advances in Neural Information Processing Systems 2017;30:6000-10.
10. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale. An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations (ICLR); 2021.
11. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. ArXiv2021. arXiv:2102.04306.
12. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. Proceedings of the European Conference on Computer Vision Workshops; 2023:205-18.
13. Gu P, Zhang Y, Wang C, Chen D. ConvFormer: Combining CNN and Transformer for Medical Image Segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). IEEE; 2023:1-5.
14. Zhang Y, Liu H, Hu Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2021. Cham: Springer International Publishing; 2021:14-24.
15. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo WY, Dollár P, Girshick R. Segment Anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023:3992-4003.
16. Roy S, Wald T, Koehler G, Rokuss MR, Disch N, Holzschuh J, Zimmerer D, Maier-Hein KH. Sam.md: Zero-shot medical image segmentation capabilities of the segment anything model. arXiv 2023; arXiv:2304.05396.
17. Baharoon M, Qureshi W, Ouyang J, Xu Y, Aljouie A, Peng W. Towards General Purpose Vision Foundation Models for Medical Image Analysis: An Experimental Study of DINOv2 on Radiology Benchmarks. arXiv 2023; arXiv:2312.02366.
18. Towle B, Chen X, Zhou K. SimSAM: Zero-Shot Medical Image Segmentation via Simulated Interaction. In: 2024 IEEE 21th International Symposium on Biomedical Imaging (ISBI). IEEE; 2024:1-5.
19. Hu C, Li X. When SAM Meets Medical Images: An Investigation of Segment Anything Model (SAM) on Multi-phase Liver Tumor Segmentation. arXiv 2023; arXiv:2304.08506.
20. Wang A, Islam M, Xu M, Zhang Y, Ren H. Sam meets robotic surgery: An empirical study on generalization, robustness and adaptation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2023. Cham: Springer International Publishing; 2023:234-44.
21. He S, Bao R, Li J, Grant PE, Ou Y. Accuracy of segment-anything model (sam) in medical image segmentation tasks. CoRR 2023;2304.09324.
22. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun 2024;15:654.
23. Cheng J, Ye J, Deng Z, Chen J, Li TX, Wang H, Su Y, Huang Z, Chen J, Jiang L, Sun H, He J, Zhang S, Zhu M, Qiao Y. SAM-Med2D. arXiv 2023; arXiv:2308.16184.
24. Wang H, Guo S, Ye J, Deng Z, Cheng J, Li TX, Chen J, Su YC, Huang Z, Shen Y, Fu B, Zhang S, He J, Qiao Y. SAM-Med3D. arXiv 2023; arXiv:2310.15161.
25. Zhao Z, Zhang Y, Wu C, Zhang X, Zhang Y, Wang Y, Xie W. One Model to Rule them All: Towards Universal Segmentation for Medical Images with Text Prompts. arXiv 2023; arXiv:2312.17183.
26. arXiv. Available online: https://arxiv.org/
27. ResearchGate. Available online: https://www.researchgate.net/
28. Google Scholar. Available online: https://scholar.google.com/
29. SEMANTIC SCHOLAR. Available online: https://www.semanticscholar.org/
30. PubMed. Available online: https://pubmed.ncbi.nlm.nih.gov/
31. OpenAI. GPT-4 Technical Report. arXiv 2023; arXiv:2303.08774.
32. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, Sutskever I. Zero-Shot Text-to-Image Generation. arXiv 2021; arXiv:2102.12092.

33. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning Transferable Visual Models From Natural Language Supervision. In: International conference on machine learning. PmLR; 2021:8748-63.
34. Wang X, Zhang X, Cao Y, Wang W, Shen C, Huang T. SegGPT: Segmenting Everything In Context. arXiv 2023; arXiv:2304.03284.
35. Zou X, Yang J, Zhang H, Li F, Li L, Gao J, Lee YJ. Segment Everything Everywhere All at Once. arXiv 2023; arXiv:2304.06718.
36. Yi H, Qin Z, Lao Q, Xu W, Jiang Z, Wang D, Zhang S, Li K. Towards General Purpose Medical AI: Continual Learning Medical Foundation Model. arXiv 2023; arXiv:2303.06580.
37. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv 2023; arXiv:2304.07193.
38. Zhang S, Metaxas D. On the challenges and perspectives of foundation models for medical image analysis. Med Image Anal 2024;91:102996.
39. Moawad A, Ahmed A, ElMohr M, Eltaher M, Habra M, Fisher S, Perrier N, Zhang M, Fuentes D, Elsayes K. Voxel-level segmentation of pathologically-proven Adrenocortical carcinoma with Ki-67 expression (Adrenal-ACC-Ki67-Seg). The Cancer Imaging Archive 2023. Available online: https://doi.org/10.7937/1FPG-VM46
40. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022:16000-9.
41. Gao Y, Zhou M, Liu D, Yan Z, Zhang S, Metaxas DN. A Data-scalable Transformer for Medical Image Segmentation: Architecture, Model Efficiency, and Benchmark. arXiv 2022; arXiv:2203.00131.
42. Koleilat T, Asgariandehkordi H, Rivaz H, Xiao Y. MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2024. Cham: Springer International Publishing; 2024:643-53.
43. Wu J, Wang Z, Hong M, Ji W, Fu H, Xu Y, Xu M, Jin Y. Medical SAM adapter: Adapting segment anything model for medical image segmentation. Med Image Anal 2025;102:103547.
44. Hu M, Li Y, Yang X. SkinSAM: adapting the segmentation anything model for skin cancer segmentation. In: Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment. SPIE; 2024:174-84.
45. Huang Z, Wang H, Deng Z, Ye J, Su Y, Sun H, He J, Gu Y, Gu L, Zhang S, Qiao Y. STU-Net: Scalable and Transferable Medical Image Segmentation Models Empowered by Large-Scale Supervised Pre-training. arXiv 2023; arXiv:2304.06716.
46. Butoi VI, Ortiz JJG, Ma T, Sabuncu MR, Guttag JV, Dalca AV. UniverSeg: Universal Medical Image Segmentation. 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023:21381-94.
47. Li Y, Hu M, Yang X. Polyp-SAM: transfer SAM for polyp segmentation. In: Medical Imaging 2024: Computer-Aided Diagnosis. SPIE; 2024:759-65.
48. Zhang K, Liu D. Customized Segment Anything Model for Medical Image Segmentation. arXiv 2023; arXiv:2304.13785.
49. Chai S, Jain RK, Teng S, Liu J, Li Y, Tateyama T, Chen YW. Ladder Fine-tuning approach for SAM integrating complementary network. arXiv 2023; arXiv:2306.12737.
50. Gong S, Zhong Y, Ma W, Li J, Wang Z, Zhang J, Heng PA, Dou Q. 3DSAM-adapter: Holistic adaptation of SAM from 2D to 3D for promptable tumor segmentation. Med Image Anal 2024;98:103324.
51. Wang G, Wu J, Luo X, Liu X, Li K, Zhang S. MIS-FM: 3D Medical Image Segmentation using Foundation Models Pretrained on a Large-Scale Unannotated Dataset. arXiv 2023; arXiv:2306.16925.
52. Shaharabany T, Dahan A, Giryes R, Wolf L. AutoSAM: Adapting SAM to Medical Images by Overloading the Prompt Encoder. arXiv 2023; arXiv:2306.06370.
53. Zhang J, Ma K, Kapse S, Saltz J, Vakalopoulou M, Prasanna P, Samaras D. Sam-path: A segment anything model for semantic segmentation in digital pathology. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2021. Cham: Springer International Publishing; 2023:161-70.
54. Shi X, Chai S, Li Y, Cheng J, Bai J, Zhao G, Chen YW. Cross-modality Attention Adapter: A Glioma Segmentation Fine-tuning Method for SAM Using Multimodal Brain MR Images. arXiv 2023; arXiv:2307.01124.
55. Holroyd NA, Li Z, Walsh CL, Brown E, Shipley RJ, Walker-Samuel S. tUbe net: a generalisable deep learning tool for 3D vessel segmentation. bioRxiv 2023; 2023.07.24.550334.
56. Yue W, Zhang J, Hu K, Xia Y, Luo J, Wang Z. SurgicalSAM: Efficient Class Promptable Surgical

Instrument Segmentation. arXiv 2023; arXiv:2308.08746.
57. Biswas R. Polyp-SAM++: Can A Text Guided SAM Perform Better for Polyp Segmentation? arXiv 2023; arXiv:2308.06623.
58. Paranjape JN, Nair NG, Sikder S, Vedula S, Patel VM. AdaptiveSAM: Towards Efficient Tuning of SAM for Surgical Scene Segmentation. In: Annual Conference on Medical Image Understanding and Analysis. Cham: Springer Nature, Switzerland; 2024:187-201.
59. Bui NT, Hoang DH, Tran MT, Le N. SAM3D: Segment Anything Model in Volumetric Medical Images. 2024 IEEE International Symposium on Biomedical Imaging (ISBI). IEEE; 2023:1-4.
60. Wang C, Chen X, Ning H, Li S. SAM-OCTA: A Fine-Tuning Strategy for Applying Foundation Model OCTA Image Segmentation Tasks. In: 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2024:1771-5.
61. Anand D, Reddy G, Singhal V, Shanbhag D, Shriram K, Patil U, Bhushan C, Manickam K, Gui D, Mullick R, Gopal A, Bhatia P, Kass-Hout TA. One-shot Localization and Segmentation of Medical Images with Foundation Models. arXiv 2023; arXiv:2310.18642.
62. Kim S, Kim K, Hu J, Chen C, Lyu Z, Hui R, Kim S, Liu Z, Zhong A, Li X, Liu T, Li Q. Medivista-sam: Zero-shot medical video analysis with spatio-temporal sam adaptation. arXiv 2023; arXiv:2309.13539.
63. Chen C, Miao J, Wu D, Zhong A, Yan Z, Kim S, Hu J, Liu Z, Sun L, Li X, Liu T, Heng PA, Li Q. MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. Med Image Anal 2024;98:103310.
64. Ravishankar H, Patil R, Melapudi V, Suthar H, Anzengruber S, Bhatia P, Taha KH, Annangi P. SonoSAMTrack--Segment and Track Anything on Ultrasound Images. arXiv 2023; arXiv:2310.16872.
65. Liu H, Yang H, Diest PJ, Pluim JPW, Veta M. WSI-SAM: Multi-resolution Segment Anything Model (SAM) for histopathology whole-slide images. arXiv 2024; arXiv:2403.09257.
66. Pandey SK, Chen KF, Dam E. Comprehensive Multimodal Segmentation in Medical Imaging: Combining YOLOv8 with SAM and HQ-SAM Models. Proceedings of the IEEE/CVF international conference on computer vision (ICCV); 2023:2592-8.
67. Sathish R, Venkataramani R, Shriram KS, Sudhakar P. Task-driven Prompt Evolution for Foundation Models. arXiv 2023; arXiv:2310.17128.
68. Ravishankar H, Patil R, Melapudi V, Annangi P. SonoSAM - Segment Anything on Ultrasound Images. In: International Workshop on Advances in Simplifying Medical Ultrasound. Cham: Springer Nature, Switzerland; 2023:23.
69. Xu Y, Tang J, Men A, Chen Q. EviPrompt: A Training-Free Evidential Prompt Generation Method for Segment Anything Model in Medical Images. arXiv 2023; arXiv:2311.06400.
70. Lei W, Xu W, Li K, Zhang X, Zhang S. MedLSAM: Localize and segment anything model for 3D CT images. Med Image Anal 2025;99:103370.
71. Israel U, Marks M, Dilip R, Li Q, Yu C, Laubscher E, Li S, Schwartz M, Pradhan E, Ates A, Abt M, Brown C, Pao E, Pearson-Goulart A, Perona P, Gkioxari G, Barnowski R, Yue Y, Valen DAV. A Foundation Model for Cell Segmentation. bioRxiv 2024; 2023.11.17.567630.
72. Du Y, Bai F, Huang T, Zhao B. SegVol: Universal and Interactive Volumetric Medical Image Segmentation. arXiv 2023; arXiv:2311.13385.
73. Xu J. GMISeg: General Medical Image Segmentation without Re-Training. arXiv 2023; arXiv:2311.12539.
74. Zhao Y, Zhou T, Gu Y, Zhou Y, Zhang Y, Wu Y, Fu H. Segment Anything Model-guided Collaborative Learning Network for Scribble-supervised Polyp Segmentation. arXiv 2023; arXiv:2312.00312.
75. Yue W, Zhang J, Hu K, Wu Q, Ge Z, Xia Y, Luo J, Wang Z. Part to Whole: Collaborative Prompting for Surgical Instrument Segmentation. arXiv 2023; arXiv:2312.14481.
76. Yan X, Sun S, Han K, Le TT, Ma H, You C, Xie X. AFTer-SAM: Adapting SAM with Axial Fusion Transformer for Medical Imaging Segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2024:7975-84.
77. Fischer M, Bartler A, Yang B. Prompt tuning for parameter-efficient medical image segmentation. Med Image Anal 2024;91:103024.
78. Gu H, Colglazier R, Dong H, Zhang J, Chen Y, Yildiz Z, et al. SegmentAnyBone: A universal model that segments any bone at any location on MRI. Med Image Anal 2025;101:103469.
79. Chen Z, Xu Q, Liu X, Yuan Y. UN-SAM: Universal Prompt-Free Segmentation for Generalized Nuclei Images. arXiv 2024; arXiv:2402.16663.
80. Xu Z, Tang F, Quan Q, Yao Q, Zhou SK. APPLE: Adversarial Privacy-aware Perturbations on Latent Embedding for Unfairness Mitigation. arXiv 2024; arXiv:2403.05114.
81. Killeen BD, Wang LJ, Zhang H, Armand M, Taylor

RH, Dreizin D, Osgood G, Unberath M. FluoroSAM: A Language-aligned Foundation Model for X-ray Image Segmentation. arXiv 2024; arXiv:2403.08059.
82. Li H, Liu H, Hu D, Wang J, Oguz I. Promise: Prompt-Driven 3D Medical Image Segmentation Using Pretrained Image Foundation Models. In: 2024 IEEE 21st International Symposium on Biomedical Imaging (ISBI). IEEE; 2023:1-5.
83. Qiu Y, Xie Z, Jiang Y, Ma J. Segment anything with inception module for automated segmentation of endometrium in ultrasound images. J Med Imaging (Bellingham) 2024;11:034504.
84. Zhou K, Yu D. Segmentation of Brain MRI Tumors by MedSAM with Prompts Generated by Object Detection. In: 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). IEEE; 2024:2081-5.
85. Wu J, Xu M, editors. One-Prompt to Segment All Medical Images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024:11302-12.
86. Indelman HC, Dahan E, Perez-Agosto AM, Shiran C, Shaked D, Daniel N. Semantic Segmentation Refiner for Ultrasound Applications with Zero-Shot Foundation Models. Annu Int Conf IEEE Eng Med Biol Soc 2024;2024:1-7.
87. Xu W, Moffat M, Seale T, Liang Z, Wagner F, Whitehouse D, Menon D, Newcombe V, Voets N, Banerjee A, Kamnitsas K. Feasibility and benefits of joint learning from MRI databases with different brain diseases and modalities for segmentation. arXiv 2024; arXiv:2405.18511.
88. Soberanis-Mukul RD, Cheng J, Mangulabnan JE, Vedula SS, Ishii M, Hager GD, Taylor RH, Unberath M. GSAM+Cutie: Text-Promptable Tool Mask Annotation for Endoscopic Video. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2024:2388-94.
89. Gao Y, Li Z, Liu D, Zhou M, Zhang S, Metaxas DN. Training Like a Medical Resident: Context-Prior Learning Toward Universal Medical Image Segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023:11194-204.
90. Cai T, Yan H, Ding K, Zhang Y, Zhou Y. WSPolyp-SAM: Weakly Supervised and Self-Guided Fine-Tuning of SAM for Colonoscopy Polyp Segmentation. Applied Sciences 2024;14:5007.
91. Zhang X, Ou NJ, Basaran BD, Visentin M, Qiao M, Gu R, Cheng O, Liu Y, Matthew PM, Ye C, Bai W. A Foundation Model for Brain Lesion Segmentation with Mixture of Modality Experts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature, Switzerland; 2024:379-89.
92. Xu Q, Li J, He X, Liu Z, Chen Z, Duan W, Li C, He MM, Tesema FB, Cheah WP, Wang Y, Qu R, Garibaldi JM. ESP-MedSAM: Efficient Self-Prompting SAM for Universal Medical Image Segmentation. arXiv 2024; arXiv:2407.14153.
93. Huo X, Tian S, Zhou B, Yu L, Li A. Dr-SAM: U-Shape Structure Segment Anything Model for Generalizable Medical Image Segmentation. In: International Conference on Intelligent Computing. Singapore: Springer Nature, Singapore; 2024:197-207.
94. Gowda SN, Clifton DA. CC-SAM: SAM with Cross-feature Attention and Context for Ultrasound Image Segmentation. arXiv 2024; arXiv:2408.00181.
95. Lin X, Xiang Y, Zhang L, Yang X, Yan Z, Yu L. SAMUS: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv 2023; arXiv:230906824.
96. Gao Y, Xia W, Hu D, Gao X. DeSAM: Decoupled Segment Anything Model for Generalizable Medical Image Segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI 2024. Cham: Springer Nature, Switzerland; 2024:509-19.
97. Yang S, Bi H, Zhang H, Sun J. SAM-UNet: Enhancing Zero-Shot Segmentation of SAM for Universal Medical Images. arXiv 2024; arXiv:2408.09886.
98. Cox J, Liu P, Stolte SE, Yang Y, Liu K, See KB, Ju H, Fang R. BrainSegFounder: Towards 3D foundation models for neuroimage segmentation. Med Image Anal 2024;97:103301.
99. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville AC, Bengio Y. Generative adversarial networks. Communications of the ACM 2020;63:139-44.
100. Hu JE, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Chen W. LoRA: Low-Rank Adaptation of Large Language Models. arXiv 2021; arXiv:2106.09685.
101. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D. Unetr: Transformers for 3d medical image segmentation. Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV); 2022:574-84.
102. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: Swin transformers for semantic segmentation

of brain tumors in MRI images. In: International MICCAI brainlesion workshop. Cham: Springer International Publishing; 2021:272-84.
103. Yan X, Tang H, Sun S, Ma H, Kong D, Xie X. After-unet: Axial fusion transformer unet for medical image segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022:3971-81.
104. Ye J, Cheng J, Chen J, Deng Z, Li TX, Wang H, Su YC, Huang Z, Chen J, Jiang L, Sun H, Zhu M, Zhang S, He J, Qiao Y. SA-Med2D-20M Dataset: Segment Anything in 2D Medical Imaging with 20 Million Masks. arXiv 2023; arXiv:2311.11969.
105. Chen Z, Deng Y, Wu Y, Gu Q, Li Y. Towards understanding mixture of experts in deep learning. arXiv 2022; arXiv:2208.02813.
106. Li Y, Sun J, Tang CK, Shum H. Lazy snapping. ACM Transactions on Graphics 2004;23:303-8.
107. Liu Q, Xu Z, Bertasius G, Niethammer M. SimpleClick: Interactive Image Segmentation with Simple Vision Transformers. Proc IEEE Int Conf Comput Vis 2023;2023:22233-43.
108. Chen X, Zhao Z, Zhang Y, Duan M, Qi D, Zhao H. FocalClick: Towards Practical Interactive Image Segmentation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022:1290-9.
109. Grady L. Random walks for image segmentation. IEEE Trans Pattern Anal Mach Intell 2006;28:1768-83.
110. Moawad A, Fuentes D, Morshid A, Khalaf A, Elmohr M, Abusaif A, Hazle J, Kaseb A, Hassan M, Mahvash A. Multimodality annotated HCC cases with and without advanced imaging segmentation. 2021. Available online: https://www.cancerimagingarchive.net/collection/hcc-tace-seg
111. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. arXiv 2020; arXiv:2005.14165.
112. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. Proc Conf Empir Methods Nat Lang Process 2022;2022:3876-87.
113. Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, Li CY, Yang J, Su H, Zhu JJ, Zhang L. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In: European Conference on Computer Vision (ECCV). Cham: Springer Nature, Switzerland; 2024:38-55.
114. Vasantharajan C, Tun KZ, Thi-Nga H, Jain S, Rong T, Siong CE. MedBERT: A Pre-trained Language Model for Biomedical Named Entity Recognition. In: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE; 2022:1482-8.
115. Zhang S, Xu Y, Usuyama N, Xu H, Bagga J, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv 2023; arXiv:2303.00915.
116. Chen P, Li Q, Biaz S, Bui T, Nguyen AT. gScoreCAM: What Objects Is CLIP Looking At? In: Proceedings of the Asian Conference on Computer Vision (ACCV); 2022:1959-75.
117. Wasserthal J, Breit HC, Meyer MT, Pradella M, Hinck D, Sauter AW, Heye T, Boll DT, Cyriac J, Yang S, Bach M, Segeroth M. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiol Artif Intell 2023;5:e230024.
118. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML); 2020:1597-607.
119. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020:9729-38.
120. Xie Z, Zhang Z, Cao Y, Lin Y, Bao J, Yao Z, Dai Q, Hu H. Simmim: A simple framework for masked image modeling. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022:9643-53.
121. Lester B, Al-Rfou R, Constant N. The Power of Scale for Parameter-Efficient Prompt Tuning. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2021:3045-59.
122. Sung YL, Cho J, Bansal M. LST: Ladder Side-Tuning for Parameter and Memory Efficient Transfer Learning. arXiv 2022; arXiv:2206.06522.
123. Jocher G, Qiu J, Chaurasia A. Ultralytics YOLO. 2023. Available online: https://ultralytics.com
124. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26:1045-57.
125. Kaggle. Available online: https://www.kaggle.com
126. Zendo. Available online: https://zenodo.org/
127. IEEE DataPort. Available online: https://ieee-

dataport.org/
128. Grand Challenge. Available online: https://grand-challenge.org/
129. Glaister J, Wong A, Clausi DA. Segmentation of skin lesions from digital images using joint statistical texture distinctiveness. IEEE Trans Biomed Eng 2014;61:1220-30.
130. Silva J, Histace A, Romain O, Dray X, Granado B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int J Comput Assist Radiol Surg 2014;9:283-93.
131. Wang Z, Lu B, Long Y, Zhong F, Cheung TH, Dou Q, Liu Y. AutoLaparo: A New Dataset of Integrated Multi-tasks for Image-guided Surgical Automation in Laparoscopic Hysterectomy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature, Switzerland; 2022:486-96.
132. Qian B, Chen H, Wang X, Guan Z, Li T, Jin Y, et al. DRAC 2022: A public benchmark for diabetic retinopathy analysis on ultra-wide optical coherence tomography angiography images. Patterns (N Y) 2024;5:100929.
133. Gupta A, Gehlot S, Goswami S, Motwani S, Gupta R, Faura ÁG, et al. SegPC-2021: A challenge & dataset on segmentation of Multiple Myeloma plasma cells from microscopic images. Med Image Anal 2023;83:102677.
134. Franco-Barranco D, Lin Z, Jang WD, Wang X, Shen Q, Yin W, et al. Current Progress and Challenges in Large-Scale 3D Mitochondria Instance Segmentation. IEEE Trans Med Imaging 2023;42:3956-71.
135. Kovalyk O, Morales-Sánchez J, Verdú-Monedero R, Sellés-Navarro I, Palazón-Cabanes A, Sancho-Gómez JL. PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. Sci Data 2022;9:291.
136. Hatamizadeh A, Hosseini H, Patel N, Choi J, Pole CC, Hoeferlin CM, Schwartz SD, Terzopoulos D. RAVIR: A Dataset and Methodology for the Semantic Segmentation and Quantitative Analysis of Retinal Arteries and Veins in Infrared Reflectance Imaging. IEEE J Biomed Health Inform 2022;26:3272-83.
137. Zhang S. Pathological images for gland segmentation. IEEE Dataport; 2023. Available online: https://dx.doi.org/10.21227/rkqj-zd61
138. Sirinukunwattana K, Pluim JPW, Chen H, Qi X, Heng PA, Guo YB, Wang LY, Matuszewski BJ, Bruni E, Sanchez U, Böhm A, Ronneberger O, Cheikh BB, Racoceanu D, Kainz P, Pfeiffer M, Urschler M, Snead DRJ, Rajpoot NM. Gland segmentation in colon histology images: The glas challenge contest. Med Image Anal 2017;35:489-502.
139. Verma R, Kumar N, Patil A, Kurian NC, Rane S, Graham S, et al. MoNuSAC2020: A Multi-Organ Nuclei Segmentation and Classification Challenge. IEEE Trans Med Imaging 2021;40:3413-23.
140. Han C, Pan X, Yan L, Lin H, Li B, Yao S, et al. WSSS4LUAD: Grand Challenge on Weakly-supervised Tissue Semantic Segmentation for Lung Adenocarcinoma. arXiv 2022; arXiv:2204.06455.
141. Chowdhury MEH, Rahman T, Khandakar AA, Mazhar R, Kadir MA, Mahbub ZB, Islam KR, Khan MS, Iqbal A, Al-Emadi NA, Reaz MBI. Can AI Help in Screening Viral and COVID-19 Pneumonia? IEEE Access 2020;8:132665-76.
142. Rahman T, Khandakar A, Qiblawey Y, Tahir A, Kiranyaz S, Abul Kashem SB, Islam MT, Al Maadeed S, Zughaier SM, Khan MS, Chowdhury MEH. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. Comput Biol Med 2021;132:104319.
143. Popov M, Amanturdieva A, Zhaksylyk N, Alkanov A, Saniyazbekov A, Aimyshev T, Ismailov E, Bulegenov A, Kuzhukeyev A, Kulanbayeva A, Kalzhanov A, Temenov N, Kolesnikov A, Sakhov O, Fazli S. Dataset for Automatic Region-based Coronary Artery Disease Diagnostics Using X-Ray Angiography Images. Sci Data 2024;11:20.
144. Chakravarty A, Sarkar T, Sathish R, Sethuraman R, Sheet D. Re-curated Breast Imaging Subset DDSM Dataset (RBIS-DDSM). IEEE Dataport; 2022. Available online: https://dx.doi.org/10.21227/nqp1-sp19
145. Gut D. X-ray images of the hip joints. Mendeley Data 2021. Available online: https://data.mendeley.com/datasets/zm6bxzhmfz/1
146. Homayoun H, Chan WY, Kuzan TY, Ling LW, Altintoprak KM, Mohammadi A, Vijayananthan A, Rahmat K, Leong SS, Mirza-Aghazadeh-Attari M, Ejtehadifar S, Faeghi F, Acharya UR, Ardakani AA. Applications of machine-learning algorithms for prediction of benign and malignant breast lesions using ultrasound radiomics signatures: A multi-center study. Biocybernetics and Biomedical Engineering 2022;42:921-33.
147. Hamyoon H, Yee Chan W, Mohammadi A, Yusuf Kuzan T, Mirza-Aghazadeh-Attari M, Leong WL, Murzoglu Altintoprak K, Vijayananthan A, Rahmat K, Ab Mumin N, Sam Leong S, Ejtehadifar S, Faeghi F, Abolghasemi J, Ciaccio EJ, Rajendra Acharya U, Abbasian Ardakani A. Artificial intelligence, BI-RADS evaluation and morphometry: A novel combination to diagnose breast

cancer using ultrasonography, results from multi-center cohorts. Eur J Radiol 2022;157:110591.
148. Rodrigues PS. Breast Ultrasound Image. Mendeley Data 2018. Available online: https://data.mendeley.com/datasets/wmy84gzngw/1
149. Da Correggio KS, Galluzzo RN, Santos LO, Barroso FSM, Chaves TZL, Onofre ASC, von Wangenheim A. Fetal Abdominal Structures Segmentation Dataset Using Ultrasonic Images. Mendeley Data 2023. Available online: https://data.mendeley.com/datasets/4gcpm9dsc3/1
150. Luo G, Xu M, Chen H, Liang X, Tao X, Ni D, et al. Tumor detection, segmentation, and classification challenge on automated 3d breast ultrasound. In: 26th International Conference on Medical Image Computing and Computer Assisted Intervention-MICCAI 2023. Available online: https://doi.org/10.5281/zenodo.2022
151. Behboodi B, Carton FX, Chabanas M, Ribaupierre SD, Solheim O, Munkvold BKR, Rivaz H, Xiao Y, Reinertsen I. RESECT-SEG: Open access annotations of intra-operative brain tumor ultrasound images. arXiv 2022; arXiv:2207.07494.
152. Quinton F, Popoff R, Presles B, Leclerc S, Mériaudeau F, Nodari G, Lopez O, Pellegrinelli J, Chevallier O, Ginhac D, Vrigneaud JM, Alberini JL. A Tumour and Liver Automatic Segmentation (ATLAS) Dataset on Contrast-Enhanced Magnetic Resonance Imaging for Hepatocellular Carcinoma. Data 2023;8:79.
153. Shapey J, Kujawa A, Dorent R, Wang G, Dimitriadis A, Grishchuk D, Paddick I, Kitchen N, Bradford R, Saeed SR, Bisdas S, Ourselin S, Vercauteren T. Segmentation of vestibular schwannoma from MRI, an open annotated dataset and baseline algorithm. Sci Data 2021;8:286.
154. Wang M, Wang H. Pituitary Adenoma MRI Segmentation Dataset. IEEE Dataport; 2023. Available online: https://dx.doi.org/10.21227/66ks-t035
155. Malinin A, Band N, Chesnokov G, Gal Y, Gales MJF, Noskov A, Ploskonosov A, Prokhorenkova L, Provilkov I, Raina V, Raina V, Shmatova M, Tigas P, Yangel B. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks. arXiv 2021; arXiv:2107.07455.
156. Malinin A, Athanasopoulos AN, Barakovic M, Cuadra MB, Gales MJF, Granziera C, Graziani M, Kartashev NK, Kyriakopoulos KG, Lu PJ, Molchanova N, Nikitakis A, Raina V, Rosa FL, Sivena E, Tsarsitalidis V, Tsompopoulou E, Volf E. Shifts 2.0: Extending The Dataset of Real Distributional Shifts. arXiv 2022; arXiv:2206.15407.
157. Gharleghi R, Adikari D, Ellenberger K, Ooi SY, Ellis C, Chen CM, et al. Automated segmentation of normal and diseased coronary arteries - The ASOCA challenge. Comput Med Imaging Graph 2022;97:102049.
158. Yao Z, Xie W, Zhang J, Dong Y, Qiu H, Yuan H, Jia Q, Wang T, Shi Y, Zhuang J, Que L, Xu X, Huang M. ImageTBAD: A 3D Computed Tomography Angiography Image Dataset for Automatic Segmentation of Type-B Aortic Dissection. Front Physiol 2021;12:732711.
159. Jonathan L, Julien O, Arnaud A. CT Training and validation series for 3D automated segmentation of inner ear using U-NET architecture deep-learning model. IEEE Dataport; 2023. Available online: https://dx.doi.org/10.21227/y91d-4p39
160. Luo X, Fu J, Zhong Y, Liu S, Han B, Astaraki M, et al. SegRap2023: A benchmark of organs-at-risk and gross tumor volume Segmentation for Radiotherapy Planning of Nasopharyngeal Carcinoma. Med Image Anal 2025;101:103447.
161. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 1945;26:297-302.
162. Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, et al. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. J Med Internet Res 2021;23:e26151.
163. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2009:248-55.
164. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. Proceedings of the IEEE international conference on computer vision (ICCV); 2017:843-52.
165. Azad B, Azad R, Eskandari S, Bozorgpour A, Kazerouni A, Rekik I, Merhof D. Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision. arXiv 2023; arXiv:2310.18689.
166. Mazurowski MA, Dong H, Gu H, Yang J, Konz N, Zhang Y. Segment anything model for medical image analysis: An experimental study. Med Image Anal 2023;89:102918.

Cite this article as: Noh S, Lee BD. A narrative review of foundation models for medical image segmentation: zero-shot performance evaluation on diverse modalities. Quant Imaging Med Surg 2025;15(6):5825-5858. doi: 10.21037/qims-2024-2826

