分割一切模型用于医学影像分析？

黄雨灏*，杨鑫*，刘恋，周涵，常澳，周心睿，陈汝锶，余俊轩，陈炯权，陈超宇，刘思菁，池昊哲，胡歆迪，岳珂娟，李雷，Vicente Grau，范登平，董发进#，倪东#

摘要

Segment Anything Model（SAM）是首个用于通用图像分割领域的基础模型，其在各种自然图像分割任务上取得了显著的表现。然而，医学图像分割往往更具挑战性，因为其涉及到复杂的目标形状、精细的解剖结构、不确定和复杂的目标边界，以及多样的目标尺度等。因此，为了充分验证 SAM 在医学影像上的性能，本文收集整理了 53 个开源数据集，并构建了一个大型医学分割数据集，其中包含 18 个模态，84 种分割目标，125 类目标-模态配对样本，共 105 万张 2D 图像和 603 万个掩模（称为 COSMOS 1050K）。本文在构建的 COSMOS 1050K 数据集上全面分析了不同的模型和测试策略。我们的发现主要包括：1) SAM 在某些特定的解剖目标上表现出了显著的性能，但在其他情况下表现不稳定甚至完全失败。2) SAM 使用 ViT-H 模型作为编码器相对于 ViT-B 模型表现更佳。3) SAM 在手动提示下，尤其是在框提示下，表现出比 Everything 模式更好的性能。4) SAM 能够有效辅助人工标注，提高标注质量，缩短标注时间。5) SAM 对中心点和外接框提示的随机偏移很敏感，可能导致性能下降。6) 在只有一个点或几个点提示时，SAM 的表现优于传统交互式方法，但随着点的数量增加，SAM 的性能会被超越。7) SAM 的性能与解剖结构的边界复杂度、前背景灰度差异等因素相关。8) 在特定的医学任务上对 SAM 进行微调可以将其基于 ViT-B 和 ViT-H 骨干网络下的平均 DICE 性能分别提高 4.39%和 6.68%。代码和模型可在：https://github.com/yuhoo0302/Segment-Anything-Model-for-Medical-Images 上获得。我们希望这份综合性报告能够帮助研究人员深入探索 SAM 在医学影像分割中的应用潜力，并为其提供正确使用和开发 SAM 的指导。

1. 引言

ChatGPT[1]和 GPT-4[2]等大语言模型的出现使得自然语言处理（Natrural Language Processing，NLP）领域进入一个新时代，这些模型展现出卓越的零样本和少样本泛化能力。这一进展激发了研究人员开发类似规模的计算机视觉（Computer Vision，CV）基础模型。最初提出的计算机视觉基础模型主要基于预训练方法，如 CLIP[1]和 ALIGN[2]。CLIP 通过将视觉概念和细节与相应的文本描述关联起来，能够识别和理解结构形状、纹理和颜色等视觉信息。这使得 CLIP 能够执行各种任务，包括图像分类、目标检测，甚至是视觉问答。ALIGN 能够生成图像区域的自然语言描述，相比传统图像描述方法，能提供更详细和可解释的结果。DALL·E[3]可用于

[1] https://chat.openai.com
[2] https://openai.com/research/gpt-4

从文本描述中生成图像。该模型在大量文本-图像对的数据集上进行训练，可以生成各种图像，从逼真的物体到结合多个概念的超现实场景。然而，这些模型并未明确针对图像分割进行优化，特别是医学图像分割的任务。

最近，Kirillov 等人创新地提出一个图像分割的基础模型，名为 Segment Anything Model（SAM）[4]。SAM 基于 Vision Transformer（ViT）[5]模型，并在包含 1100 万图像和 10 亿掩模的大型数据集上进行了训练。SAM 最大的亮点是其对未知数据集和任务的良好零样本分割性能。这一过程由不同的提示（例如点和框）驱动，用于指示目标结构的像素级语义和区域级位置信息。SAM 已被证实是高度通用的，其能够处理各种不同的分割任务[4]。

基于 SAM 的预训练模型，一些论文进一步研究了其在不同零样本分割场景中的性能。我们粗略地将它们分为两类：1）非医学图像上的应用和 2）医学图像上的应用。

1.1 SAM 在非医学图像应用中

两项研究专注于测试 SAM 在 Everything 模式下在分割伪装目标方面的性能[6], [7]。结果显示，在这些场景中（例如在自然环境中视觉上隐藏的伪装动物），其性能较差。同时，作者发现 SAM 在工业场景中无法检测隐藏的缺陷。随后，Ji 等人探讨了 SAM 的三种测试策略（点、框和 Everything）在各种应用中的效果[8]。具体而言，他们的任务涵盖了常规自然图像（显著/伪装/透明目标分割和阴影检测）、农业（作物分割和害虫与叶病监测）、制造业（异常和表面缺陷检测）以及遥感（建筑和道路提取）。他们得出结论，尽管 SAM 在某些场景中可以取得良好的性能，比如显著目标分割和农业分析，但在其他应用中产生了较差的结果。他们还验证了与 Everything 模式相比，人工提示可以有效地改善分割结果。

1.2 SAM 在医学图像分析中

Ji 等人评估了 SAM 在 Everything 模式下对不同解剖结构（例如，脑、肺和肝脏）和模态（计算机断层扫描 CT 和磁共振成像 MRI）的分割表现[7]。实验结果表明，SAM 在分割具有清晰边界的器官时表现较好，但在识别边界模糊病变区域时可能会有困难。另一项研究评估了 SAM 在一些医疗领域的性能（视盘视杯、息肉和皮肤病变分割），采用了 Everything 和两种人工提示（点和框）策略。作者发现 SAM 需要相当多的人类先验知识（即点）才能在这些任务上获得相对准确的结果；否则，SAM 可能导致错误的分割，尤其是在没有提示的情况下[8]。在 MRI 大脑提取任务中，Mohapatra 等人将 SAM 与 FMRIB 软件库的大脑提取工具（Brain Extraction Tool，BET）进行了比较[9]。定量结果显示，SAM 的分割结果优于 BET，说明 SAM 在大脑提取任务中的应用中有发展的潜力。Deng 等人评估了 SAM 在数字病理学分割任务中的性能，包括在整片成像上进行的肿瘤、非肿瘤组织和细胞核分割[10]。结果表明，SAM 对大型连通目标的分割结果出色。然而，对于密集实例目标分割，即使用所有目标框或每张图像 20 个点的提示，SAM 依然无法实现令人满意的性能。Zhou 等人采用了 Everything 分割策略将 SAM 应用于五个基准数据集的息肉分割任务[11]。结果表明，尽管 SAM 在某些情况下可以准确分割息肉，但 SAM 与最先进方法之间存在较大差距。此外，Liu 等人将 3D

Slicer 软件与 SAM 结合，以协助 SAM 在医学图像的开发、评估和利用[12]。

最近，几项研究在大于等于 10 个公共医学影像分割数据集或任务上测试了 SAM 的性能。在 He 等人的研究中，得出了 SAM 的零样本分割性能明显低于传统基于深度学习方法的结论[13]。在 Mazurowski 等人的研究中，作者使用不同数量的点提示评估了 SAM 的性能[14]。他们观察到随着点数的增加，SAM 的性能会趋于稳定。此外，他们发现 SAM 的表现： 1）总体分割性能一般，2）在不同数据集和任务之间性能不稳定。Ma 和 Wang 验证了原始 SAM 在许多医学数据集上可能会完全失败，平均 DICE 指标只有 58.52%[15]。然后，他们使用医学图像对 SAM 进行了微调，所提出的 MedSAM 相比原始 SAM 在 DICE 上提高了 22.51%。Wu 等人采用了适应器 Adapter 技术对 SAM 进行微调，并增强了其医学目标感知能力[16]。实验证实，他们提出的医学 SAM Adapter 可以胜过最先进的医学图像分割方法（如 nnUnet 等[17]）。

尽管以上研究调查了 SAM 在医学图像分割中的性能，但它们至少存在以下一种缺点：

1. 数据集较小。以往的研究主要侧重于对 SAM 在 MRI、CT 和数字病理学等模态中的性能进行评估，其中仅包含了有限数量的分割目标。然而，医学图像通常涵盖多个模态，并且需要对许多解剖结构或其他目标进行分割。这限制了先前的研究在医学影像分割领域的综合分析[7]–[11]。

2. 单一的 SAM 测试策略。大多数以往的研究仅采用有限，甚至只有一种类型的测试策略对 SAM 进行评估[7], [11], [14]。然而，不同的医学目标通常表现出不同的特征，因此可能需要针对它们自身的特点设计相应的测试模式。有限的测试策略可能导致对 SAM 的分析不够准确和完整。

3. 缺乏全面深入的评估。一些现有的研究[8]仅通过在线演示提供的可视化结果对 SAM 进行评估[3]。此外，一些研究仅使用有限的度量指标（如 DICE 或 IOU）来评估 SAM 的性能[10]；大多数研究没有探索 SAM 对医学目标的感知能力。因此，SAM 的分割性能与医学分割目标属性之间的关系尚未经过详尽研究[13]–[15]。

对医学分割目标感知的分析非常重要，它可以帮助社区更好地了解影响 SAM 分割性能的因素（即感知医学目标的能力），以更好地开发和改进新一代通用医学分割模型。在这个研究中，本文建立了一个名为 COSMOS 1050K 的大型医学图像数据集，包括 1050K 张 2D 图像，涵盖了 18 种不同的模态（见图 1）和 84 个分割目标（例如解剖结构、病变、细胞、工具等），且覆盖整个人体（见图 2）。这可以帮助我们全面分析和评估 SAM 在医学图像上的性能。然后，本文充分探讨 SAM 的不同测试策略，并提供丰富的定量和定性实验结果，展示 SAM 对医学目标的感知能力。最后，本文深入评估了 SAM 的性能与目标属性（例如边界复杂性、前景背景对比度和目标大小等）之间的关系。我们希望这份全面的研究能为社区提供对医学 SAM 未来发展的一些见解。

[3] https://segment-anything.com/demo

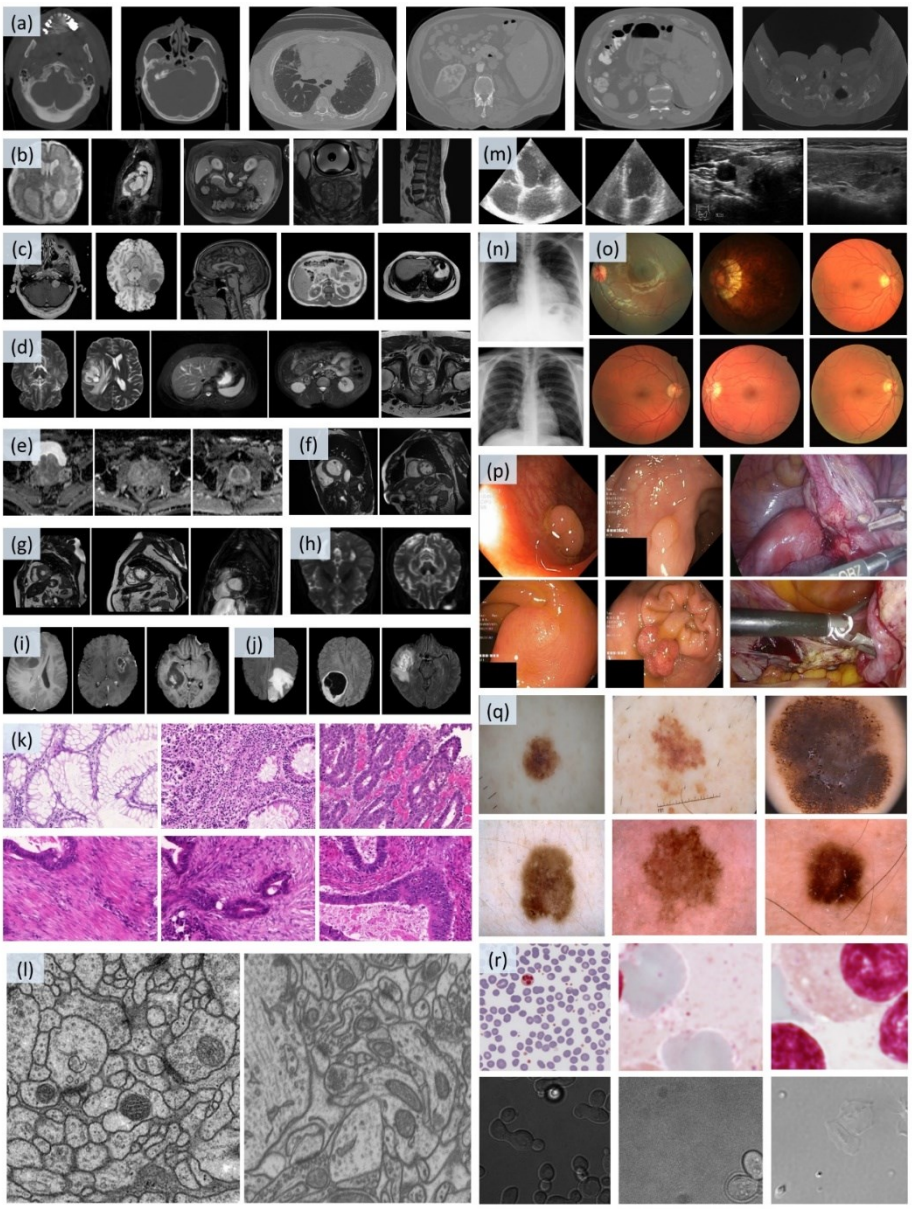


图 1. COSMOS 1050K 数据集涵盖了各种模态，包括(a) CT，(b) MRI，(c) T1W MRI，(d) T2W MRI，(e) ADC MRI，(f) Cine-MRI，(g) CMR，(h) DW MRI，(i) T1-GD MRI，(j) T2-FLAIR MRI，(k) 组织病理学，(l) 电子显微镜，(m) 超声，(n) X 射线，(o) 眼底，(p) 结肠镜，(q) 皮肤镜，以及(r) 显微镜。

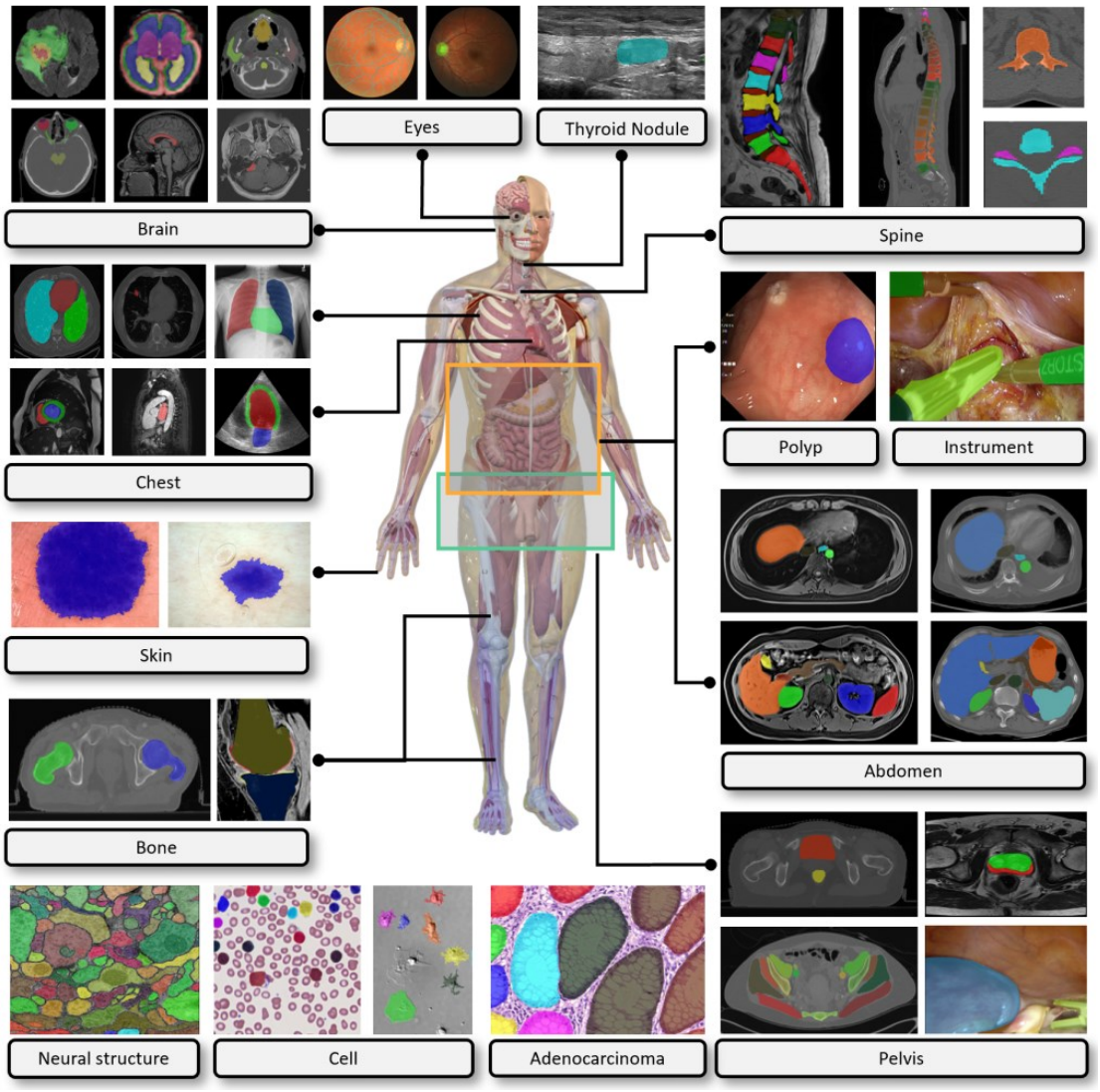
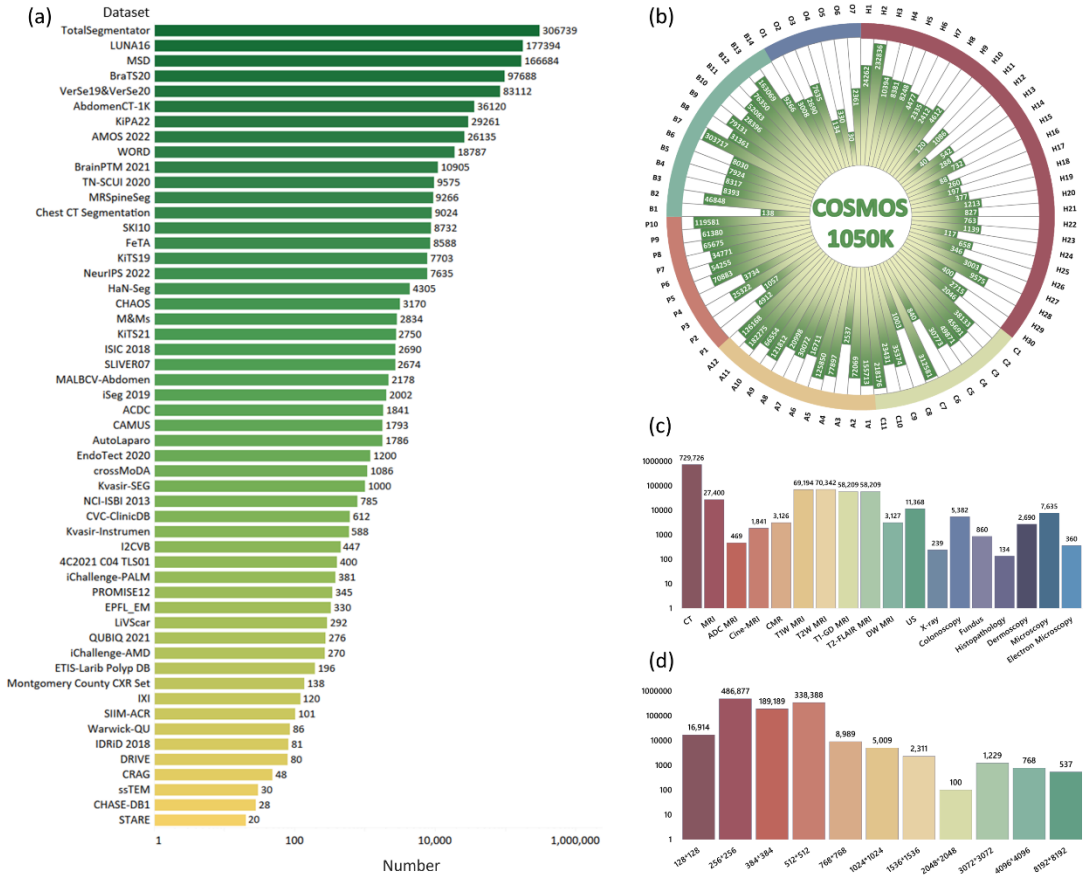


图 2. COSMOS 1050K 数据集涵盖了覆盖全身的大多数生物医学目标，例如脑肿瘤、眼底血管、甲状腺结节、脊柱、肺部、心脏、腹部器官和肿瘤、细胞、息肉和手术器械。

2. 数据集

医学图像具有多种模态，如 CT、MRI、超声和 X 射线等。不同模态之间存在很大的域差异[18]，且各种模态在可视化特定目标（包括解剖结构和病变）方面都具有优势[12]。为了充分评估 SAM 在医学图像分割中的泛化性能，本文收集了 53 个公共数据集，并对它们进行标准化，构建了大型 COSMOS 1050K 数据集。对于 COSMOS 1050K 中的模态分类汇总方式，本文参考了每个公共数据集的官方介绍和最近发表的研究[19]（更多详情请见表 1）。图 1 和图 2 分别展示了数据集中的各种成像模态和有代表性的分割目标。本文将在以下两个方面详细描述 COSMOS 1050K 数据集，包括图像收集和预处理规范。



H1: Brain	**H22:** Submandibular Gland	**A2:** Pancreas	**B1:** Arytenoids
H2: Brain Tumor	**H23:** Oral Cavity	**A3:** Pancreatic Mass	**B2:** Head of Femur
H3: Cerebrospinal Fluid	**H24:** Pituitary Gland	**A4:** Spleen	**B3:** Femur
H4: Brain Grey Matter	**H25:** Larynx Supraglottic	**A5:** Kidney	**B4:** Femur Cartilage
H5: Brain White Matter	**H26:** Larynx Glottis	**A6:** Kidney Tumor	**B5:** Tibia
H6: Brain Ventricles	**H27:** Carotid Artery	**A7:** Adrenal Gland	**B6:** Tibia Cartilage
H7: Cerebellum	**H28:** Thyroid Nodules	**A8:** Gall Bladder	**B7:** Vertebra
H8: Brain Deep Grey Matter	**H29:** Throat Cancer	**A9:** Esophagus	**B8:** Clavicula
H9: Brainstem	**H30:** Spinal Cord	**A10:** Stomach	**B9:** Hip
H10: Callosum	**C1:** Heart	**A11:** Arota	**B10:** Humerus
H11: Vestibular Schwannomas	**C2:** Left Ventricle	**A12:** Inferior Vena Cava	**B11:** Sacrum
H12: Eyeball	**C3:** Left Ventricular Myocardium	**P1:** Prostate	**B12:** Scapula
H13: Posterior Eyeball	**C4:** Right Ventricle	**P2:** Uterus	**B13:** Rib
H14: Anterior Eyeball	**C5:** Left Atrium	**P3:** Bladder	**B14:** Spine
H15: Optic Disc	**C6:** Left Atrium Appendage	**P4:** Rectum	**O1:** Polyp
H16: Fundus Vasculature	**C7:** Lung	**P5:** Gluteus Maximus	**O2:** Melanoma
H17: Lacrimal Gland	**C8:** Lung Tumor	**P6:** Gluteus Medius	**O3:** Cell
H18: Optic Nerve	**C9:** Right Atrium	**P7:** Gluteus Minimus	**O4:** Adenocarcinoma
H19: Cricopharyngeal	**C10:** Pulmonary Artery	**P8:** Iliac Artery	**O5:** Mitochondria
H20: Parotid Gland	**C11:** Sacrospinalis	**P9:** Iliac Vena	**O6:** Neural Structures
H21: Lips	**A1:** Liver	**P10:** Iliopsoas	**O7:** Tool

图 3. COSMOS 1050K 的统计信息。(a)收集到的公开数据集处理后的数据量；(b)84 种分割目标的直方图分布；(c)18 种影像模态的直方图分布；(d)图像分辨率的直方图分布。

表 1. COSMOS 1050K 数据集介绍。

Dataset Name	Description	Image Modalities
AbdomenCT-1K (Ma et al., 2021)	Liver, kidney, spleen and pancreas	CT
ACDC (Bernard et al., 2018)	Left and right ventricle and left ventricular myocardium	Cine-MRI
AMOS 2022 (Ji et al., 2022)	Abdominal multi-organ segmentation	CT, MRI
AutoLaparo (Wang et al., 2022)	Integrated dataset with multiple image-based perception tasks	Colonoscopy
BrainPTM 2021 (Avital et al., 2019; Nelkenbaum et al., 2020)	white matter tracts	T1W MRI, DW MRI
BraTS20 (Menze et al., 2014; Bakas et al., 2017, 2018)	Brain tumor	T1W MRI, T2W MRI, T1-GD MRI, T2-FLAIR MRI
CAMUS (Leclerc et al., 2019)	Four-chamber and Apical two-chamber heart	US
CellSeg Challenge-NeurIPS 2022 (Ma et al., 2023)	Cell segmentation	Microscopy
CHAOS (Kavur et al., 2021)	Livers, kidneys and spleens	CT, T1W MRI, T2W MRI
CHASE-DB1 (Zhang et al., 2016)	Retinal vessel segmentation	Fundus
Chest CT Segmentation (Dataset, a)	Lungs, heart and trachea	CT
CRAG (Graham et al., 2019)	Colorectal adenocarcinoma	Histopathology
crossMoDA (Shapey et al., 2019, 2021)	Vestibular schwannoma	T1W MRI
CVC-ClinicDB (Bernal et al., 2015)	Polyp	Colonoscopy
DRIVE (Liu et al., 2022b)	Retinal vessel segmentation	Fundus
EndoTect 2020 (Hicks et al., 2021)	Polyp	Colonoscopy
EPFL-EM (Lucchi et al., 2013)	Mitochondria and synapses segmentation	Electron Microscopy
ETIS-Larib Polyp DB (Bernal et al., 2017; Yoon et al., 2022)	Polyp	Colonoscopy
FeTA (Payette et al., 2021)	Seven tissues of the infant brain	T2W MRI
HaN-Seg (Podobnik et al., 2023)	Healthy organs-at-risk near the head and neck	CT
I2CVB (Lemaître et al., 2015)	Prostate	T2W MRI
iChallenge-AMD (Li et al., 2020)	Optic disc and fovea	Fundus
iChallenge-PALM (Huazhu et al., 2019)	Optic disc and lesions from pathological myopia patients	Fundus
IDRiD 2018 (Porwal et al., 2018)	Optic disc, fovea and lesion segmentation	Fundus
iSeg 2019 (Sun et al., 2021)	White matter, gray matter, and cerebrospinal fluid of infant brain	T1W MRI, T2W MRI
ISIC 2018 (Tschandl et al., 2018; Codella et al., 2018, 2019)	Melanoma of skin	Dermoscopy
IXI (Dataset, c)	Callosum	T1W MRI
KiPA22 (He et al., 2021, 2020; Shao et al., 2011, 2012)	Kidney, tumor, renal vein and renal artery	CT
KiTS19 (Heller et al., 2020)	Kidneys and tumors	CT
KiTS21 (Zhao et al., 2022)	Kidneys, cysts, tumors, ureters, arteries and veins	CT
Kvasir-Instrumen (Jha et al., 2021)	Gastrointestinal procedure instruments such as snares, balloons, etc.	Colonoscopy
Kvasir-SEG (Jha et al., 2020)	Gastrointestinal polyp	Colonoscopy
LiVScar (Karim et al., 2016)	Infarct segmentation in the left ventricle	CMR
LUNA16 (Setio et al., 2017)	Lungs, heart and trachea	CT
M&Ms (Campello et al., 2021)	Left and right ventricle and left ventricular myocardium	CMR
MALBCV-Abdomen (Landman et al., 2015)	Abdominal multi-organ segmentation	CT
Montgomery County CXR Set (Jaeger et al., 2014)	Lung	X-ray
MRSpineSeg (Pang et al., 2020)	multi-class segmentation of vertebrae and intervertebral discs	MRI
MSD (Antonelli et al., 2022; Simpson et al., 2019)	Large-scale collection of 10 Medical Segmentation Datasets	CT, MRI, ADC MRI, T1W MRI, T2W MRI, T1-GD MRI, T2-FLAIR MRI
NCI-ISBI 2013 (Li et al., 2013)	Prostate (peripheral zone, central gland)	T2W MRI
PROMISE12 (Litjens et al., 2014)	Prostate	T2W MRI
QUBIQ 2021 (Ji et al., 2021)	Kidney, prostate, brain growth, and brain tumor	CT, MRI, T1W MRI, T2W MRI, T1-GD MRI, T1-FLAIR MRI
SIIM-ACR (Zawacki et al., 2019; Viniavskyi et al., 2020)	Pneumothorax segmentation	X-ray
SKI10 (Lee et al., 2010)	Cartilage and bone segmentation from knee data	MRI
SLIVER07 (Heimann et al., 2009)	Liver	CT
ssTEM (Cardona et al., 2010)	Neuronal structures	Electron Microscopy
STARE (Hoover et al., 2000; Hoover and Goldbaum, 2003)	Retinal vessel segmentation	Fundus
TN-SCUI 2020 (Zhou et al., 2020)	Thyroid nodule	US
TotalSegmentator (Wasserthal et al., 2023)	Multiple anatomic structures segmentation(27 organs, 59 bones, 10 muscles, and 8 vessels)	CT
VerSe19 (Sekuboyina et al., 2021)	Spine or vertebral segmentation	CT
VerSe20 (Löffler et al., 2020; Liebl et al., 2021)	Spine or vertebral segmentation	CT
Warwick-QU (Sirinukunwattana et al., 2017)	Gland segmentation	Histopathology
WORD (Luo et al., 2022)	Abdominal multi-organ segmentation	CT
4C2021 C04 TLS01 (Dataset, b)	Throat and hypopharynx cancer lesion area	CT

2.1 数据集收集

医学图像涵盖了多种目标结构类型，如脑器官和肿瘤[20]–[24]，肺和心脏[25]–[28]，腹部[20], [29]–[32]，脊柱[33]–[35]，细胞[36], [37]和息肉[38], [39]等等。表 1 详细列出了收集的医学图像分割数据集，图 3（a）展示了每个数据集在预处理后的图像数量。为了兼容不同的 SAM 评估模式，本文采用了以下排除标准：1）排除极小的结构，如图 4（a）中所示的耳蜗和输尿管。这是因为在极小目标上自动生成点或框提示的难度较大。2）排除在三维数据在切片时，在二维切面上会明显分离的结构，如肠道（图 4（b）），下颌和甲状腺。其目的是避免产生提示混淆。3）排除整体结构相对分散的目标，如乳腺癌的组织学图像（见图 4（c）），肺气管树切片（见图 4（d）），肾动脉和静脉。这些结构大多分散在 2D 切片中，并会嵌入在周围的组织/结构中，导致在这些结构上无法合理地使用 SAM 的提示模式进行验证。根据上述标准，COSMOS 1050K 现在包括总共 84 个目标，它们的数量如图 3（b）所示。在统计信息时，部分目标结构在一张图像中仅被分类统计一次，不区分位置或详细的类别划分（例如，形状相似的“左肺”和“右肺”被归为“肺”，各种无明确类别信息的手术工具被归为“工具”）。图 3 下方的图例中有更多详细信息。图 3（c）和图 3（d）分别显示了模态和图像分辨率的直方图分布。鉴于相同目标在不同模态之间存在显著差异，包括灰度分布和纹理特征的差异，本文将它们进一步分为 125 个目标-模态配对的目标（例如肝脏包含 CT 和 MRI 等模态：Liver-CT 和 Liver-MRI）。

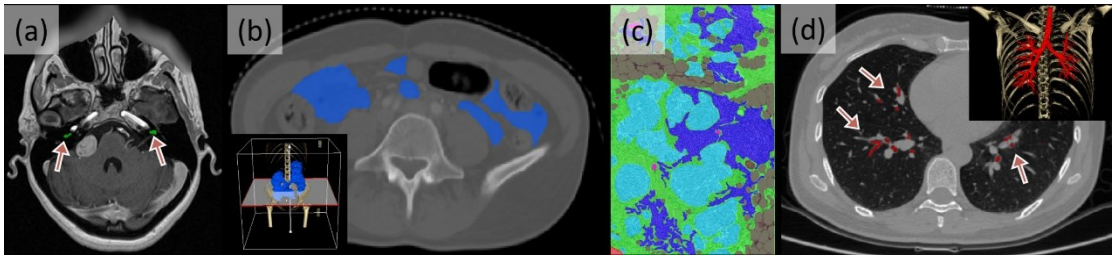


图 4. 符合排除标准的典型例子。 (a) 耳蜗（标准 1），(b) 肠道（标准 2），(c) 组织病理学乳腺癌（标准 3），以及(d) 肺气管树（标准 3）。图中(b)和(d)的角落显示了通过 Pair 注释软件包获得的 3D 渲染图像。

2.2 数据集预处理规范

COSMOS 1050K 包含不同的标签、模态、格式和形状。此外，SAM 的原始版本仅支持 2D 输入，且 2D 格式是 3D/4D 格式的基本组成部分。为了在不同数据集之间标准化数据，我们对每个收集的公共数据集应用了以下预处理步骤。

对于 3D 容积，预处理过程可以总结如下：1）提取沿主视图平面的切片，因为它的分辨率较高。在 CT 中，通常是横断面，而在 MRI 中，可能是横断面，例如前列腺、脑肿瘤，或矢状面，例如脊柱和心脏。2）保留标签像素总和大于 50 的切片，以确保每个切片包含相应的正确标签。3）通过最小-最大归一化来标准化图像：$I_n = 255 * (I - I_{min}) / (I_{max} - I_{mim})$，取值范围为（0，255）。I 表示原始提取的图像，I_n表示标准化后的图像。I_{min} 和 I_{max} 分别是 I 的最小和最大灰度值。这是因为医学图像的体素或像素值可能在很大程度上变化。例如，MRI 的强度范围为（0，800），CT 的强度范围为（-2000，2000），而其他模态可能已经在范围（0，255）内[19]。同时，我们根据目标结构的类别或位置（例如，左肾和右肾具有不同的像素值）重新设置掩模的像素值（1-255）。4）以 PNG 格式保存图像和标签。对于 4D 数据（N，W，H，D），我们将数据转换为 N 组 3D 体积，然后按照 3D 体积的处理流程进行处理。这里，N 表示 4D 数据中配对体积的数量。对于 2D 图像，预处理步骤如下：1）保留标签像素面积大于 50 的图像。2）根据目标的类别或位置，将标签的像素值重新设置在 1 到 255 的范围内。对于 CellSeg Challenge-NeurIPS 2022 数据集[37]，由于原始标签值的范围较广（1-1600），我们将每个图像和标签重构为多个子图，以确保统一的标签范围。3）将图像和标签的格式从 BMP、JPG、TIF 等转换为 PNG，以实现一致的数据加载。

总体而言，COSMOS 1050K 包括 1,050,311 张 2D 图像或切片，其中 1,003,809 张来自 8,653 个 3D 体积，46,502 张是独立的 2D 图像，数据集总共包含 6,033,198 个掩模。

3. 方法

3.1 SAM 介绍

SAM 与传统分割框架有所不同，它创新地提出了可提示的分割任务，该任务由灵活的、提示驱动的模型架构和广泛而多样的训练数据支持。SAM 提出了数据引擎模式，即建立一个循环数据标注-模型训练过程，利用模型促进数据收集，然后利用新收集的数据来提升模

型的性能。最终，SAM 是在一个庞大的数据集上进行训练的，该数据集包含来自 1100 万个有使用许可的 2D 图像的超过十亿个掩膜。

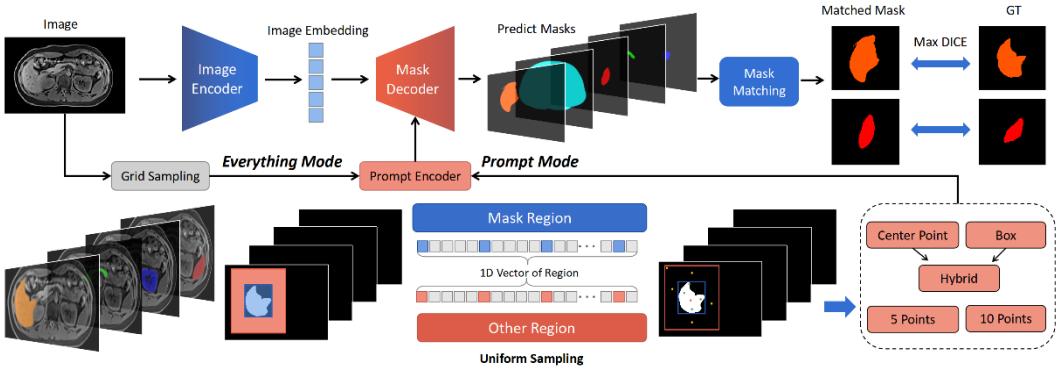


图 5. 本研究中 SAM 的测试流程。

如图 5 所示，SAM 主要包括三个组件：图像编码器（Image Encoder）、提示编码器（Prompt Encoder）和掩膜解码器（Mask Encoder）。图像编码器使用 ViT 作为骨干，在 Masked Autoencoder（MAE[40]）技术的预训练下，它以一张图像为输入，并输出图像的编码，以便与提示的编码组合。提示编码器由密集（掩膜）和稀疏（点、框和文本）分支组成。密集分支通过卷积神经网络（Convolutional Neural Network，CNN）对掩膜提示进行编码。对于稀疏分支，可以通过位置编码[41]表示点和框，而文本则由 CLIP 模型[1]进行编码。最后，所有的编码输入到掩膜解码器中，输出预测的掩膜。

在测试阶段，SAM 支持 Everything 和 Prompt 两种模式。对于前者，用户只需将一张图像输入 SAM，即可生成所有潜在目标的预测掩膜。对于后者，用户需要手动向 SAM 提供一些额外的提示，包括掩膜、框、点和文本，以向 SAM 提供有关分割目标的更多信息。这两种模式的细节将在接下来的子章节中介绍。需要注意的是，SAM 只能在图像中找到多个目标，而无法输出它们的类别（即单标签：目标或非目标结构）。

在官方的 GitHub 代码库中，SAM 的作者提供了三种不同骨干大小的预训练模型，根据模型参数从小到大，分别命名为 ViT-B、ViT-L 和 ViT-H。他们发现 ViT-H 相比于 ViT-B 在性能上表现出显著的提高，但由于其增加的复杂性，ViT-H 测试时间也成倍的增加[4]。

在评估 SAM 在医学图像中的性能时，一项研究使用了六个医学数据集，并发现 ViT-B、ViT-L 和 ViT-H 的性能没有明显的区别[42]。在本研究中，我们选择对比最小的 ViT-B（具有 12 个 Transformer 层和 91M 参数）和最大的 ViT-H（具有 32 个 Transformer 层和 636M 参数），并在大规模的 COSMOS 1050K 数据集上对这两种模型进行更广泛的验证。

3.2. Everything 全自动模式

在 Everything 模式（S_1）中，SAM 自动生成整个图像中所有潜在目标结构的分割掩膜，无需任何手动先验。该过程的初始步骤涉及生成覆盖整个图像的网格点提示（即网格采样）。基于均匀采样的网格点，提示编码器生成网格点的编码，并将其与图像的编码结合。然后，掩膜解码器将以此组合作为输入，并输出整个图像的多个潜在掩膜。然后，SAM 采用了一种

过滤机制，通过使用置信度分数、基于阈值抖动的稳定性评估以及非最大抑制（NMS）技术，从生成的掩膜中去除重复度高和低质量的部分。

3.3. Prompt 人工提示模式

在 Prompt 模式中，SAM 提供了不同类型的提示，包括点、框和文本。点提示包括正样本和负样本的点，分别表示分割目标的前景和背景。框提示用外接框表示需要分割目标的区域。此外，文本提示是使用一个句子（即关于位置、颜色、大小等方面的基本信息）来描述分割的目标。值得注意的是，文本提示的相关代码目前还未在官方的 GitHub 代码库[4]中发布。

如图 5 所示，本文设置的 Prompt 模式包含五种策略，包括一个前景点（S_2）、五个前景点（S_3）、五个前景点和五个背景点（S_4）、一个框（S_5）以及一个框和一个前景点（S_6）。本文进一步建立了一个统一的选点规则，以确保随机性、可重复性和准确性。对于前景点的选择，a）我们首先计算了真实标签（Ground Truth， GT）的质心（图 5 中的红点）。b）如果质心在 GT 掩膜内部，我们将质心作为第一个前景点。c）然后，我们将 GT 掩膜直接展平为一维向量，并采用均匀采样方法（图 5 中的绿点）获得其他前景点。d) 如果质心在 GT 掩膜外部，则通过执行步骤 c 获取所有所需的前景点。对于背景点的选择，我们希望避免选择距离目标区域 GT 太远的点。具体而言，我们首先将 GT 的边界框扩大两倍。背景点是通过在非 GT 区域进行均匀采样而生成的（图 5 中的黄点）。最后，对于框的选择，我们直接采用 GT 掩膜的最小外接正框，没有任何额外的操作。上述策略可以确保实验的可重复性。此外，我们倾向于通过选择质心和紧凑的边界框来测试 SAM 的理论最优性能。因为它们可能包含目标的最具代表性的特征。需要注意的是，SAM 允许一次将多个提示输入到网络中。因此，为了公平比较，本文在上述五种提示策略（S_2-S_6）下测试 SAM 的单轮交互性能（无特殊说明的情况）。

3.4. 推理效率

由上，我们对一张图像需要使用不同的策略进行多次测试（如 n 次），以获得最终的评估（见图 5）。然而，在 SAM 的原始代码逻辑和设计中，若对一张图像进行多策略（n 个）测试，相同的编码操作需要重复执行 n 次，这会导致多测试策略测试的运行效率较差。当使用更高分辨率输入时，推理的效率会更慢。基于这一观察，我们预先编码了所有输入图像的特征，并将其保存为中间文件。因此，基于这样的方式，只需要运行一次编码过程，便可以重复使用编码后存储的图像特征结果，以减少推断过程中的计算量。最终，SAM 测试的整体效率可以提高近 n 倍。此外，SAM 中测试的策略越多，可以节省的时间就越多。这一简单的方法可以扩展到 SAM 的其他多策略测试场景。

[4] https://github.com/facebookresearch/segment-anything

3.5. 用于分割评估的掩膜匹配机制

对每个输入图像，SAM 都会生成多个二值掩膜，但并非所有的掩膜都包含目标结构。同时，掩膜结构不包含类别信息。因此，我们提出了一种掩膜匹配机制，以评估 SAM 在每种模式下的分割性能理论上限。具体而言，对于输入图像中的一个目标 GT（前景之一），我们计算了模型输出的 N 个二值掩膜 $\{\boldsymbol{P}_n\}_{n=1}^{N}$和 GT (**G**)之间的 Dice 分数 $\{DICE_n\}_{n=1}^{N}$。然后，从该集合中选择 Dice 分数最高的掩膜，作为该目标匹配的预测掩膜 **P**，用于后续的分割评估。获得 **P** 的过程可以表达如下：

$$\boldsymbol{P} = \max\{(\mathrm{P}_1 \cdot \mathrm{G}), (P_2 \cdot G), \ldots, (P_N \cdot G)\} \tag{1}$$

其中，*N* 是一张图像中一个目标的预测二值掩膜的总数。操作（·）表示计算一个预测掩膜与 GT 之间的 Dice 分数，而 max{}表示获取具有最高 Dice 分数的预测掩膜。

4. 实验与结果

4.1. 实现细节

(1) 代码实现和逻辑。在本研究中，本文基本按照官方的 GitHub 代码库 [4] 实现了 SAM 的测试流程。对于多策略测试场景，我们需要运行 SAM 算法 n 次，并提取图像的编码 n 次。我们观察到提取图像编码的过程非常耗时，而相同的图像编码可以在不同的测试策略中重复使用，因此，我们试图优化和加速这个多次提取的过程，并对代码的部分进行了重构。对于每张测试图像，我们使用图像编码器只进行一次特征提取，并将其编码特征保存为一个 npy 文件。当应用不同的测试策略时，只需加载相应的 npy 文件，这显著提高了测试效率（约为 n×）。此外，在 Prompt 模式中，我们在图像编码后仅计算一次所需的点和框的编码结果，并将它们存储为 npz 文件。因此，所有提示测试策略可以直接使用 npz 信息，无需重新计算。

(2) 软件包的版本和功能。我们在测试中使用了多个 GPU，包括 12GB 显存的 NVIDIA GTX 2080Ti，24GB 显存的 NVIDIA GTX 3090，以及 48GB 显存的 NVIDIA A40。我们使用了 python（版本 3.8.0）、PyTorch（版本 2.0.0）和 torchvision（版本 0.15.1）来运行 SAM。我们使用了以下函数：1）使用 torch.compile 以 max-autotune 模式对模型进行封装；2）使用 torch.cuda.amp 来自适应地调整张量的类型为 Float16 或 Float32；3）使用@torch.inference 模式替代常见的 torch.no_grad，以减少 GPU 内存占用并提高推理速度，同时保持模型计算的精度。我们使用 Numpy 包（版本 1.24.2）生成/计算提示（框/点），以模拟人与测试图像的交互过程（点击、绘制框等）。此外，我们使用 OpenCV（版本 4.7.0.72）和 Matplotlib（版本 3.7.1）软件包来可视化提供的提示 prompt 和最终的分割结果。

(3) 测试策略设计。我们设计了不同的设置，充分探索 SAM 在各种测试策略下的性能。首先考虑到 SAM 的亮点功能，即 Everything 模式。它可以在没有任何手动提示的情况下输出给定图像的所有预测掩膜。考虑到医学图像分割是一项极具挑战性的任务，我们逐步引入了各种手动提示，以帮助实现准确的分割。手动提示包括：1）一个前景点，2）五个前景点，3）

五个前景点和五个背景点，4）一个框，以及 5）一个框和一个前景点。手动提供的提示可以更好地引导 SAM 输出准确的分割结果。所有的测试策略及其缩写如下所示：

- **S_1：Everything 模式；**
- **S_2：一个前景点；**
- **S_3：五个前景点；**
- **S_4：五个前景点和五个背景点；**
- **S_5：一个框；**
- **S_6：一个框和一个前景点；**

4.2. 评估指标

为了充分评估 SAM 的分割性能，我们使用了三个常见的指标，如下所示：

1. **DICE 系数**（DICE，%）：用于评估预测掩膜和 GT 之间的重叠的相似性度量。取值范围为 [0, 1]，数值越高表示模型性能越好。

2. **杰卡德相似系数**（Jaccard Similarity Coefficient，JAC，%）：也称为 IOU，用于衡量两个掩膜之间的相似性。它类似于 DICE，但具有不同的计算方法。具体而言，对于预测掩膜（A）和 GT 掩膜（B），JAC 计算交集（$|A \cap B|$）除以并集（$|A \cup B|$）。JAC 的取值范围为 0 到 1，数值越高表示性能越好。

3. **豪斯多夫距离**（Hausdorff Distance，HD，pixel）：评估两组轮廓点之间的相似度，可以反映预测掩膜中每个点到 GT 中的点的距离。与 DICE 相比，HD 对边缘相似度更敏感。

在接下来的实验中，我们主要从 DICE 和 HD 两个方面来分析实验结果。另一个相似度测量指标（即 JAC）的结果放到了补充材料中。

4.3. 不同模型下的分割性能

在本节中，我们比较了两个模型（ViT-B 和 ViT-H）在不同策略下的分割性能。从图 6 中可以观察到，在 Everything 模式（S_1）下，ViT-H 在 DICE 上比 ViT-B 高 7.47%，在 HD 上比 ViT-B 低 10.61 像素。对于一个点的提示（S_2），ViT-H 的平均性能略高于 ViT-B。随着点提示数量的增加，ViT-H 的优势将变得更加明显。而对于其余的策略（一个框/一个框+一个正样本点，S_5-S_6），它们的性能非常接近（DICE 的差异分别为 0.37%和 0.06%）。与点提示相比，框提示包含更多关于结构的区域信息。因此，它可以更好地引导 SAM 与不同模型实现更好的分割性能。具体结构的 DICE 和 HD 性能可以在表 2 和表 3 中找到。我们在正文中仅呈现了 SAM 在 40 个目标的详细分割性能，完整结果可以在附录中找到。

表 2. 不同模态下对于常见医学目标的分割 DICE（%）。 ViT-B 和 ViT-H 分别代表 SAM 的小型和大型编码器。S_1 - S_6代表不同的测试策略，包括 Everything，1 个、5 个、10 个点提示，外接框，和外接框加 1 个点提示。

Object-Modality	ViT-B						ViT-H					
	S_1	S_2	S_3	S_4	S_5	S_6	S_1	S_2	S_3	S_4	S_5	S_6
Arota-CT	57.87	68.71	71.92	74.88	86.88	85.57	65.71	70.59	76.68	81.11	86.80	85.09
Brain-CT	54.79	80.71	82.11	83.49	91.34	91.37	72.70	83.03	85.59	86.45	91.14	90.91
Heart-CT	13.01	57.23	63.53	66.65	90.15	90.50	22.33	62.10	70.63	76.75	90.06	90.06
Humerus-CT	83.54	91.66	91.93	91.91	95.16	95.07	86.61	91.93	92.46	93.31	95.28	95.17
Kidney-CT	75.27	86.86	87.28	87.84	93.60	93.26	82.88	87.66	89.49	90.88	93.29	93.09
Kidney Tumor-CT	26.83	60.57	62.98	68.93	90.69	90.63	43.22	67.30	80.27	83.23	90.74	90.95
Left Atrium-CT	13.90	34.60	35.46	44.93	87.50	87.82	22.84	36.94	40.74	55.71	87.23	86.76
Left Ventricular Myocardium-CT	7.79	27.59	29.33	37.12	63.21	63.16	13.43	27.97	32.77	42.52	63.52	63.57
Liver-CT	37.56	68.32	69.67	68.89	89.11	88.28	47.11	70.14	77.26	82.13	89.00	88.74
Lung-CT	81.36	86.05	94.45	93.64	96.75	93.89	89.35	89.59	95.65	95.75	96.73	96.40
Lung Tumor-CT	37.19	69.33	71.13	75.40	84.21	85.02	42.00	64.86	73.29	78.44	84.15	84.87
Pancreas-CT	12.75	45.88	46.56	51.81	76.86	76.38	18.30	39.59	53.37	65.77	75.86	75.59
Rib-CT	50.52	74.23	72.23	74.97	88.21	86.76	58.15	74.07	75.08	79.50	86.70	85.80
Right Ventricle-CT	13.22	36.66	36.97	47.34	82.56	81.67	20.26	38.03	42.49	57.26	82.45	81.56
Spleen-CT	39.11	70.78	72.24	76.82	92.54	92.35	50.16	73.31	78.36	82.30	92.28	92.16
Stomach-CT	27.04	52.09	54.87	61.87	83.91	84.41	34.58	54.46	65.40	74.34	83.96	84.89
Vertebra-CT	43.22	61.53	67.10	69.64	80.03	76.85	55.32	61.40	69.82	73.53	79.35	76.63
Arota-MRI	75.49	81.44	81.09	84.39	90.86	90.30	80.86	83.28	84.83	88.38	90.55	90.25
Femur-MRI	71.23	92.60	93.47	92.47	95.18	94.94	82.02	92.45	93.64	93.71	94.95	94.86
Gall Bladder-MRI	39.84	66.98	65.51	72.68	87.97	87.23	52.64	68.64	73.24	80.12	87.49	87.05
Kidney-MRI	82.30	88.83	88.72	89.72	93.65	93.28	87.19	88.79	90.05	91.30	93.26	93.06
Liver-MRI	50.32	82.33	82.48	84.07	90.72	90.42	71.72	87.24	89.16	89.91	91.40	91.31
Pancreas-MRI	15.02	49.31	48.84	58.85	79.73	79.63	27.49	50.79	61.97	71.71	79.23	78.94
Prostate-MRI	19.09	74.69	76.90	81.32	92.85	92.27	40.60	74.66	77.93	83.66	92.11	91.18
Spine-MRI	35.73	64.82	70.37	75.04	80.32	80.61	39.68	63.79	74.58	77.41	81.04	81.96
Spleen-MRI	55.43	81.00	81.56	86.62	93.74	93.24	67.80	82.66	85.60	87.94	93.04	92.68
Stomach-MRI	22.63	54.05	56.00	63.72	82.62	82.71	29.71	52.76	64.35	73.24	81.86	82.22
Tibia-MRI	80.66	93.95	94.32	93.75	96.11	95.98	88.56	94.01	94.69	94.76	96.25	96.13
Brain-T1W MRI	96.44	96.62	99.35	98.62	99.65	99.49	98.77	98.48	99.36	99.39	99.57	99.54
Brain Tumor-T1W MRI	22.22	38.45	39.09	48.68	72.29	71.96	25.53	40.24	47.29	55.49	71.50	71.92
Brain Tumor-T2W MRI	24.16	47.59	48.94	57.78	75.67	75.33	30.71	49.38	57.92	64.09	75.46	75.59
Prostate-ADC MRI	19.47	49.26	50.44	57.03	77.03	76.53	34.45	46.62	51.21	59.03	75.00	74.80
Left Ventricle-Cine-MRI	53.19	82.25	92.76	88.62	92.45	93.07	63.99	80.66	92.87	91.32	92.22	92.36
Right Ventricle-CMR	36.34	76.42	75.68	77.03	89.72	89.41	51.33	73.18	79.74	83.35	89.06	88.66
Brain-DW MRI	40.41	84.32	88.90	86.74	91.62	90.97	77.68	86.07	88.29	89.13	91.34	91.42
Brain Tumor-T1-GD MRI	26.36	41.18	43.96	52.02	71.76	71.16	32.73	42.94	50.68	57.11	70.80	70.83
Brain Tumor-T2-FLAIR MRI	25.00	51.81	52.42	61.59	77.74	77.42	35.61	56.44	64.73	69.65	77.99	78.08
Thyroid Nodules-US	31.52	66.57	76.80	78.54	90.12	90.30	48.56	71.52	80.71	83.95	89.49	89.84
Lung-X-ray	9.56	93.25	94.03	93.23	95.32	95.11	64.42	91.96	94.11	94.20	95.62	95.65
Eye-Fundus	65.14	99.30	99.19	99.08	99.15	99.22	99.22	99.24	99.26	99.23	99.31	99.28
Tool-Colonoscopy	45.59	80.93	87.27	86.02	90.89	90.89	75.55	83.49	92.31	91.58	91.47	91.62
Polyp-Colonoscopy	49.49	81.63	85.63	85.28	89.94	90.68	71.29	85.97	90.28	91.34	90.97	91.87
Adenocarcinoma-Histopathology	41.41	74.31	85.03	84.96	91.40	90.73	75.26	86.15	90.65	90.89	93.29	93.17
Melanoma-Dermoscopy	47.43	76.06	81.52	81.20	87.47	87.56	61.26	76.84	81.69	81.91	86.67	86.68
Cell-Microscopy	55.94	84.34	91.31	80.64	91.76	91.60	75.45	79.66	81.62	82.21	85.10	85.03
Neural Structures-Electron Microscopy	54.63	78.57	78.86	77.91	86.99	86.20	61.54	79.14	80.85	81.25	87.70	87.07

表 3. 不同模态下对于常见医学目标的分割 HD（像素）。 ViT-B 和 ViT-H 分别代表 SAM 的小型和大型编码器。S_1 - S_6代表不同的测试策略，包括 Everything，1 个、5 个、10 个点提示，外接框，和外接框加 1 个点提示。

Object-Modality	ViT-B						ViT-H					
	S_1	S_2	S_3	S_4	S_5	S_6	S_1	S_2	S_3	S_4	S_5	S_6
Arota-CT	47.70	21.57	18.68	17.03	7.50	7.17	39.39	20.18	13.62	10.92	7.68	7.28
Brain-CT	37.49	16.54	15.96	16.46	8.80	8.64	23.34	14.50	12.99	12.86	8.47	8.71
Heart-CT	233.23	154.20	135.36	124.77	23.49	23.21	227.05	124.27	108.98	92.50	23.73	23.67
Humerus-CT	16.41	4.77	4.95	5.42	2.24	2.37	11.99	4.86	4.06	3.84	2.19	2.21
Kidney-CT	40.48	18.47	18.43	18.94	9.65	10.25	29.20	18.71	15.44	13.36	10.37	10.52
Kidney Tumor-CT	84.17	47.15	45.58	39.83	8.10	8.47	69.88	36.69	20.62	17.90	7.86	7.83
Left Atrium-CT	81.14	48.98	49.74	44.15	4.65	5.22	71.31	42.63	40.23	29.97	4.66	5.00
Left Ventricular Myocardium-CT	105.43	60.31	57.95	42.37	12.58	12.65	99.41	56.98	48.41	36.98	12.00	12.21
Liver-CT	149.00	85.65	78.72	84.59	28.22	29.52	127.53	80.90	58.79	50.77	27.38	27.92
Lung-CT	57.69	45.90	21.56	42.70	12.11	16.76	48.44	39.36	14.98	22.63	10.76	10.50
Lung Tumor-CT	108.83	34.87	34.73	28.32	9.28	9.11	96.89	43.34	30.83	22.82	9.14	9.07
Pancreas-CT	135.29	52.93	51.52	38.48	12.39	12.63	120.75	68.38	41.53	23.39	13.48	13.50
Rib-CT	60.27	15.81	15.25	10.69	1.58	1.67	48.03	16.75	10.66	7.36	1.69	1.72
Right Ventricle-CT	101.42	61.21	59.00	36.99	9.17	9.57	93.08	53.03	44.63	31.36	9.32	9.80
Spleen-CT	125.28	46.89	42.79	34.44	6.44	6.61	100.26	46.52	33.48	24.36	6.37	6.41
Stomach-CT	93.98	59.79	59.24	41.01	11.80	11.90	84.03	57.43	41.88	26.79	11.88	11.68
Vertebra-CT	46.89	20.71	16.33	15.67	7.99	8.68	31.85	21.45	14.58	12.37	8.25	8.76
Arota-MRI	26.19	11.79	11.93	8.29	3.42	3.57	17.85	11.21	7.56	5.12	3.49	3.61
Femur-MRI	65.49	20.03	19.62	41.44	13.93	14.89	46.92	18.33	16.89	20.80	13.91	13.31
Gall Bladder-MRI	109.02	27.99	30.96	18.59	5.06	5.22	70.11	26.42	17.67	10.81	5.15	5.24
Kidney-MRI	28.27	13.43	13.36	12.55	7.67	8.72	18.35	13.69	11.57	10.41	8.38	8.46
Liver-MRI	113.93	46.71	43.86	51.74	22.66	23.54	70.66	37.84	31.01	29.66	23.17	23.28
Pancreas-MRI	123.47	48.35	50.86	32.53	11.13	11.26	93.84	43.83	25.86	18.08	11.62	11.68
Prostate-MRI	195.89	59.22	54.08	55.25	17.87	19.04	158.55	50.40	43.87	39.60	17.10	17.52
Spine-MRI	116.20	44.59	29.91	21.17	13.67	13.32	111.57	46.92	26.02	19.66	13.46	12.28
Spleen-MRI	90.79	28.53	27.65	17.61	5.87	6.25	61.74	23.78	17.26	14.83	6.06	6.32
Stomach-MRI	126.55	52.28	56.81	38.46	12.72	12.89	107.99	53.84	37.97	25.21	13.35	13.41
Tibia-MRI	49.40	14.14	19.30	48.64	7.64	8.41	27.62	11.32	10.65	14.27	6.76	6.88
Brain-T1W MRI	7.16	7.26	3.51	8.70	2.80	3.13	4.90	5.96	3.66	3.30	2.62	2.70
Brain Tumor-T1W MRI	78.08	60.05	62.43	46.91	15.35	16.13	74.02	55.87	47.83	35.79	15.16	14.83
Brain Tumor-T2W MRI	75.45	49.85	50.90	35.79	13.61	14.07	67.80	46.41	37.09	27.97	13.85	13.83
Prostate-ADC MRI	89.48	43.56	43.26	37.06	12.85	13.26	76.12	45.17	39.73	31.00	13.83	13.93
Left Ventricle-Cine-MRI	33.68	6.68	3.39	7.37	3.23	3.14	24.80	6.61	3.18	3.97	3.21	3.27
Right Ventricle-CMR	78.26	18.12	19.76	21.19	4.85	5.08	63.36	20.81	12.98	11.05	4.96	5.10
Brain-DW MRI	49.27	26.48	18.97	24.72	13.88	15.81	28.20	23.32	19.66	17.83	14.42	14.48
Brain Tumor-T1-GD MRI	69.26	51.83	54.06	41.86	15.07	15.77	60.97	48.60	41.69	31.85	15.05	15.10
Brain Tumor-T2-FLAIR MRI	75.20	46.69	48.16	33.41	13.02	13.45	63.10	39.36	30.26	23.24	13.06	13.08
Thyroid Nodules-US	163.70	77.62	63.03	62.41	20.86	21.61	127.29	63.00	42.05	34.78	20.71	20.53
Lung-X-ray	1977.68	257.39	236.67	448.30	159.31	175.42	829.27	272.97	210.19	251.03	129.96	126.08
Eye-Fundus	147.05	84.68	7.75	94.83	4.56	37.85	4.28	19.03	5.47	4.26	3.86	3.97
Tool-Colonoscopy	326.06	150.49	102.04	169.69	49.88	52.50	189.25	128.34	43.01	63.74	43.45	43.77
Polyp-Colonoscopy	214.44	111.48	101.82	110.13	59.50	60.51	145.66	96.13	78.36	78.91	50.53	50.15
Adenocarcinoma-Histopathology	299.05	89.15	63.26	93.00	30.33	32.58	126.99	55.09	41.77	46.82	23.41	24.51
Melanoma-Dermoscopy	947.16	444.84	417.83	538.32	274.31	283.00	697.27	397.32	359.48	352.98	259.81	262.04
Cell-Microscopy	82.76	17.00	13.30	28.95	6.93	7.67	33.55	24.61	22.84	23.29	16.29	16.37
Neural Structures-Electron Microscopy	126.14	20.55	18.80	30.54	7.52	8.40	95.13	19.65	14.19	19.14	6.97	7.22

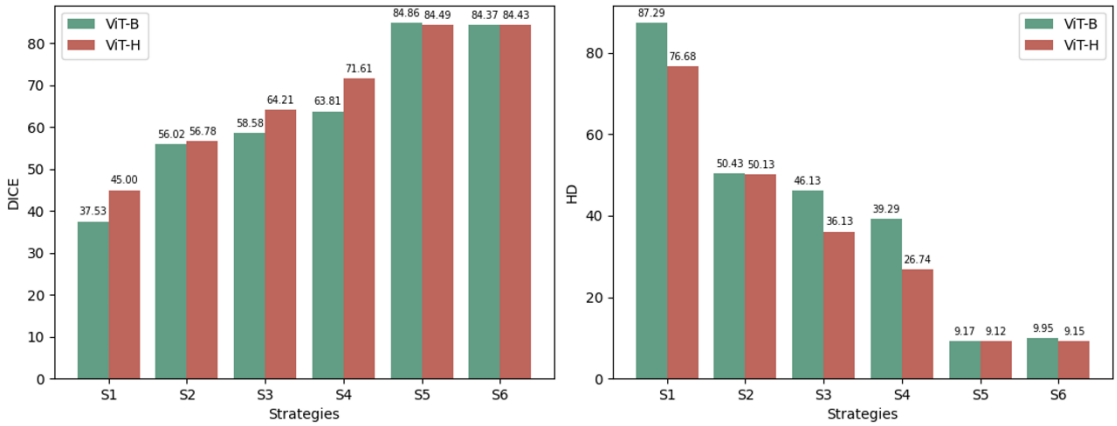


图 6. ViT-B 和 ViT-H 在不同策略下平均性能的比较。

图 7 和图 8 显示了在不同模型尺寸下评估的相同分割目标的 DICE 分布，可以证明，与 ViT-B 相比，ViT-H 下分割 DICE 分布的标准差更小，表现出更稳定的性能（以 Kidney-CT、Prostate-MRI 和 Polyp-Colonoscopy 为典型示例）。图 9 展示了分割的可视化结果。附录中还可以找到其他对比以及可视化的结果。

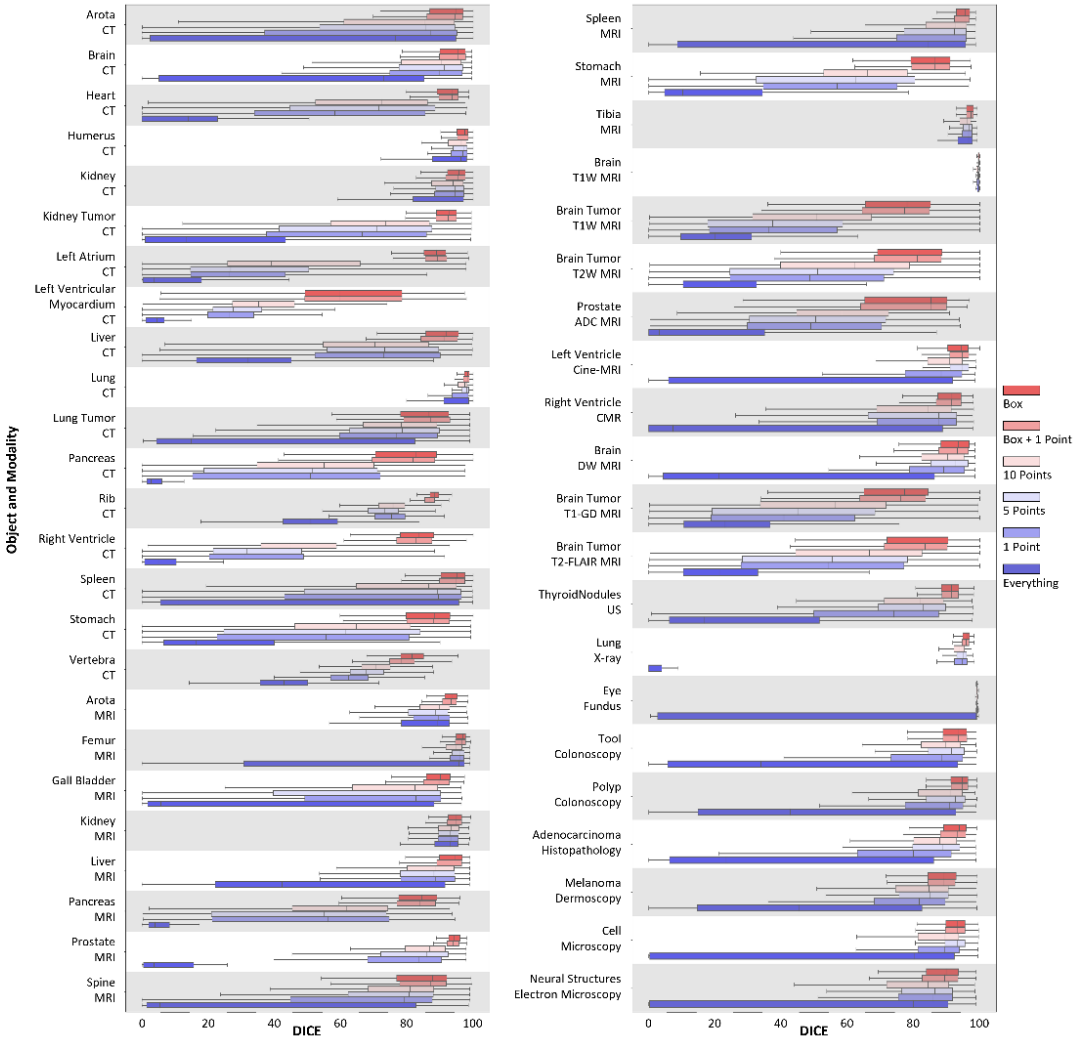


图 7. 在不同测试策略下，常见的医学目标在 ViT-B 下的 DICE 性能比较。

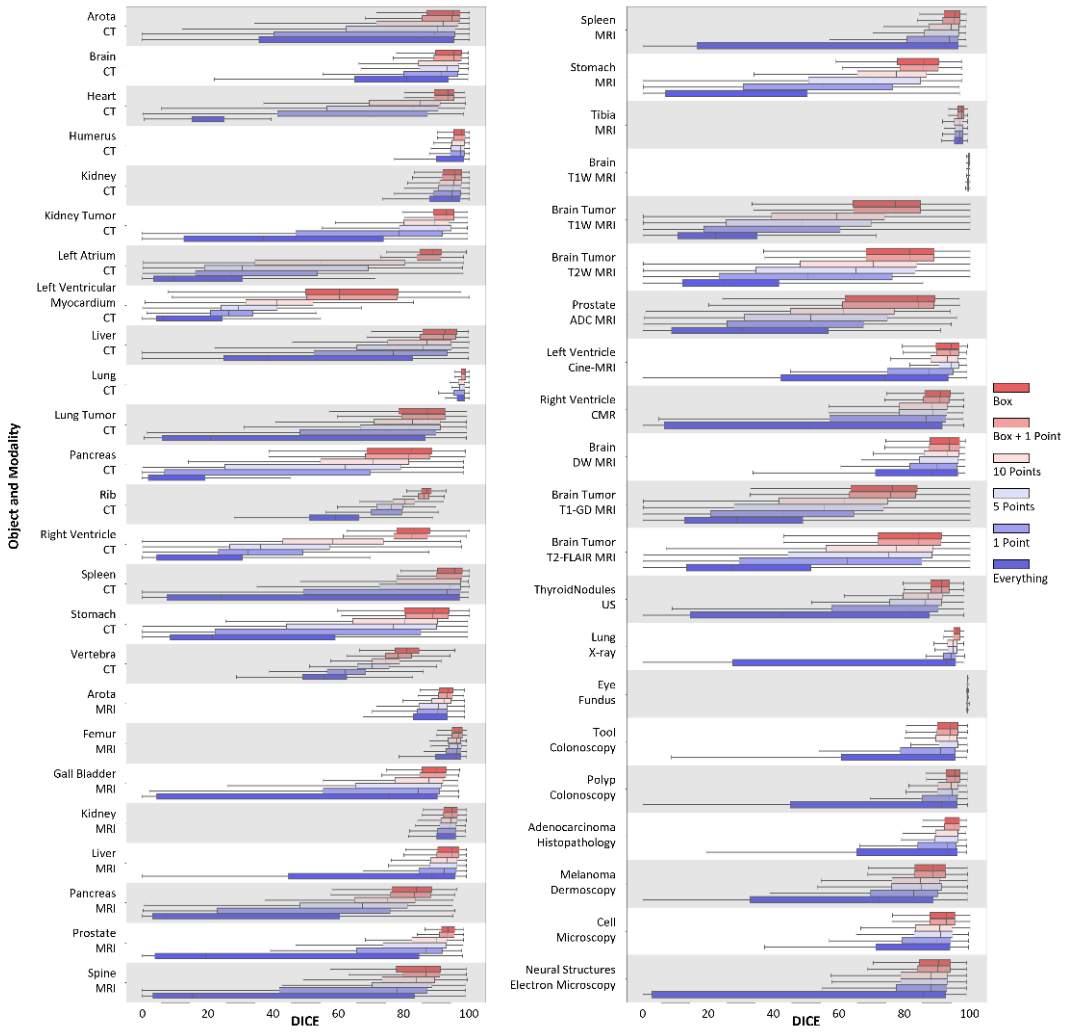


图 8. 在不同测试策略下，常见的医学目标在 ViT-H 下的 DICE 性能比较。

4.4. 在不同测试模式下的分割性能

在本节中，我们分别比较了使用不同模型（ViT-B 和 ViT-H）中不同策略之间的分割性能。如图 6 所示，我们展示了 ViT-B 和 ViT-H 在不同策略下的平均 DICE 和 HD 性能。对于 ViT-B 和 ViT-H，不同策略的性能趋势基本一致。Everything 模式（S_1）的性能是最差的。对于点提示（S_2-S_4），添加更多点将会带来稳定的性能提升（ViT-B：DICE 从 56.02%提高到 63.81%，ViT-H：DICE 从 56.78%提高到 71.61%）。使用框提示的 SAM 表现出最佳性能，而在框中添加一个前景点提示不会带来明显的变化（ViT-B：DICE 下降 0.49%，ViT-H：DICE 下降 0.06%）。

基于以上实验，我们得出结论：与点提示相比，框提示包含更多关键信息。因为框实际上指示了目标的确切位置，也给出了有限框区域范围内的潜在特征。然而，点仅表示目标的部分特征，这可能会导致模型识别和预测混淆。图 7、图 8、表 2 和表 3 呈现了在 6 种测试策略下部分目标的具体分割定量结果。图 9 展示了 5 种 Prompt 模式（S_2-S_6）预测的可视化结果。更多定量和定性结果请参阅附录。

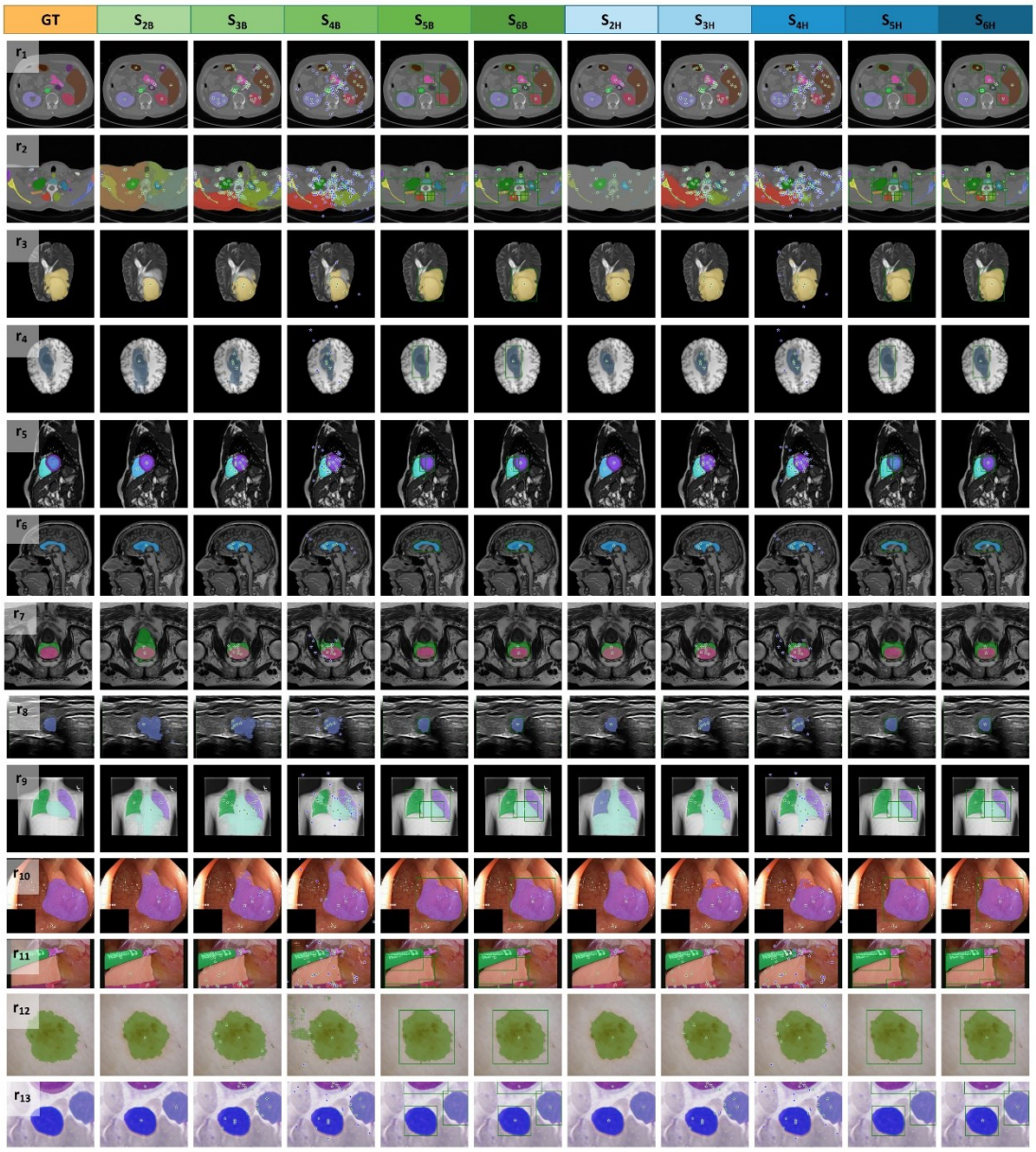


图 9. SAM 的典型良好案例（r 代表行）。 r1，r2：CT，r3，r7：T2W MRI，r4，r6：T1W MRI，r5：CMR，r8：超声，r9：X 射线，r10，r11：结肠镜，r12：皮肤镜，r13：显微镜。 绿色和蓝色星星分别表示前景和背景的点提示，绿色框表示框提示。

4.5. 不同模态医学图像的分割性能

在图 10 中，基于框提示，我们总结了 SAM 在不同模态下的性能。无论是 ViT-B 还是 ViT-H，SAM 在 X 射线模态下都有着最高的平均 DICE 性能。另外，在组织病理学、结肠镜和 DW MRI 模态中，SAM 也达到了令人满意的 DICE 性能（>90%）。此外，它们的标准差相对较低。这证明了 SAM 对这些模态的目标结构有着良好且稳定的分割性能。有 6 个（ViT-B）和 7 个（ViT-H）模态的平均 DICE 性能在 80 到 90 之间，但它们大部分有较大的标准差。SAM 在剩余的模态中的 DICE 表现较差，结果也较不稳定。请注意，SAM 对于不同模态的相同结构的分割性能会略有不同。例如，表 2 和表 3 中的 BrainTumor，其 DICE 分布从 71.76%到 77.74%，HD 分布从 13.02 像素到 15.35 像素。尽管除模态因素外，分割性能可能受到其他各种因素

的影响，但我们希望在研究 SAM 在不同模态中的使用时，图 10 的结果能为研究人员提供基本的指导。

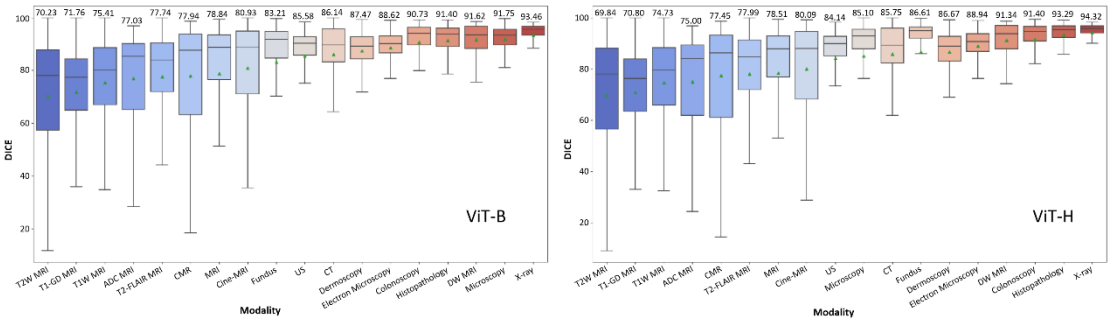


图 10. 18 种不同模态的 DICE。箱线图中的绿色三角形和上方的值是 DICE 平均值。

4.6. SAM 的推理时间分析

推理时间是评估模型的重要因素。在表 4 中，我们展示了图像编码（Embedding Generation）、提示编码（Prompt Encoding）和掩膜解码（Mask Decoding）方面的平均推理时间。所有测试都在具有 24G 内存的 NVIDIA GTX 3090 GPU 上进行。测试时间可能受到许多因素的影响，包括图像大小（在预处理时间上存在微小差异，即将具有不同尺寸的图像上采样到 1024×1024 大小）和需要分割的目标数量（以串行方式处理每个目标的提示）。因此，我们将图像大小限制为 256×256，并将分割目标数量设置为 1，以进行公平的比较。可以观察到，ViT-H 的图像编码时间几乎是 ViT-B 的四倍。Everything 模式（S_1）的 Prompt Encoding 和 Mask Decoding 过程非常耗时，因为它需要处理从整个图像采样的数百个点，包括使用 NMS 等进行大量的后处理操作。而对于手动 Prompt Encoding 和 Mask Decoding（S_2-S_6），不同模型和策略的参考时间相近且小于 0.01 秒。我们相信 SAM 在 Prompt 模式下的推理时间可以满足实时使用的需求。

表 4. SAM 的测试时间（秒）分析。

Model	Embedding	Prompt Encoding+Mask Decoding					
		S_1	S_2	S_3	S_4	S_5	S_6
ViT-B	0.1276	1.9692	0.0085	0.0088	0.0090	0.0080	0.0088
ViT-H	0.4718	3.0324	0.0086	0.0088	0.0091	0.0080	0.0090

4.7. 关于 Everything 模式中点的数量的分析

如上所述，在 Everything 模式中，将生成一个点提示的网格（$m \times m$）。默认情况下，m 设置为 32。点的数量将对最终的分割性能产生影响，特别是对于具有不同尺寸的多个目标的图像，不当的参数设计将导致分割不完整，即存在其中一些目标未获得点提示的情况。如表 5 所示，我们在四个具有多个结构的数据集上进行了测试。结果显示，在这四个数据集中，随着点的数量从 82 增加到 2562，DICE 也随之增加。图 11 还显示了更多的点将带来更多的潜在目标（不同颜色表示不同目标结构）。此外，太多的点会导致 SAM 将一个目标分割成几个部分，破坏了目标的完整性。

表 5. 在 Everything 模式下对点提示数量的消融实验。

Objects	8^2	16^2	32^2	64^2	128^2	256^2
Adenocarcinoma-Histopathology (Sirinukunwattana et al., 2017)	$42.3_{44.1}$	$70.4_{37.4}$	$76.3_{32.5}$	$77.8_{31.1}$	$78.7_{30.4}$	$79.1_{30.2}$
Adenocarcinoma-Histopathology (Graham et al., 2019)	$55.6_{42.4}$	$70.9_{34.7}$	$74.2_{31.5}$	$76.3_{30.0}$	$77.1_{29.1}$	$77.6_{28.6}$
Mitochondria-Electron Microscopy (Lucchi et al., 2013)	$31.8_{39.3}$	$71.1_{31.4}$	$81.1_{20.1}$	$81.5_{19.1}$	$81.7_{18.6}$	$81.7_{18.3}$
Neural Structures-Microscopy (Cardona et al., 2010)	$21.3_{38.0}$	$45.7_{44.1}$	$61.5_{40.1}$	$64.4_{38.4}$	$65.4_{37.8}$	$65.8_{37.5}$

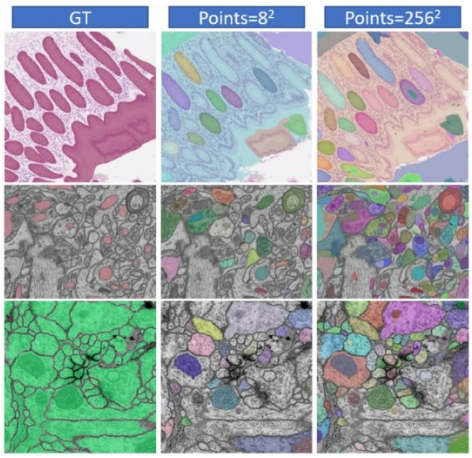


图 11. 在S_{1H}中，具有不同点提示数量的腺癌，线粒体和神经结构的不同案例。

4.8. 影响分割结果的因素分析

为验证影响 SAM 分割性能的因素，我们记录了 191,779 个解剖结构的目标属性，包括大小、长宽比、前背景灰度差异、模态以及边界复杂性。通过分析这些因素，我们旨在更好地了解解剖结构特征与 SAM 分割性能之间的相关性，并进一步为医学 SAM 的发展提供一些有用的见解。

解剖结构的大小被计算为相应掩膜的像素面积。为了确定掩膜的长宽比，我们需要计算其边界框短边和长边之间的比率（范围从 0 到 1）。灰度差异被定义为目标和扩展边界框（除目标外）背景区域的平均灰度差。具体而言，为了适应目标的各种尺寸，我们通过设定的 0.1 比率动态向外扩展边界框，而不是使用固定的像素值（例如，向外扩展 10 像素）。此外，每个解剖结构的模态被映射为数值。最后，本文引入了椭圆傅里叶描述子（Elliptical Fourier Descriptors，EFD）来描述边界复杂性。

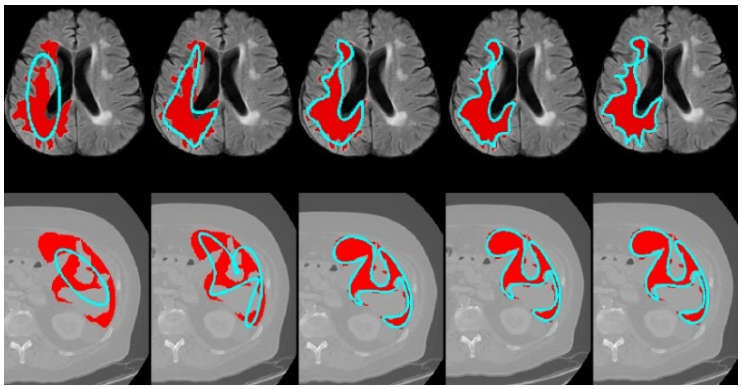


图 12. 从傅里叶级数解码的轮廓。从左到右，随着傅里叶级数的增加，解码的轮廓（蓝色）逐渐接近原始轮廓（红色）。

EFD 将掩膜的轮廓编码成代表不同频率分量的傅里叶级数。随着傅里叶阶数（Fourier order, FO）的增加，从傅里叶级数解码出的轮廓越来越接近原始轮廓（见图 12），解码过程可以描述为公式 2 和公式 3。

$$x_N(t) = L_x + \sum_{n=1}^{N}\left(a_n sin\left(\frac{T}{2n\pi t}\right) + b_n cos\left(\frac{T}{2n\pi t}\right)\right) \quad (2)$$

$$y_N(t) = L_y + \sum_{n=1}^{N}\left(c_n sin\left(\frac{T}{2n\pi t}\right) + d_n cos\left(\frac{T}{2n\pi t}\right)\right) \quad (3)$$

其中，$\left(x_N(t), y_N(t)\right)$ 是轮廓上任意点的坐标，N 是傅里叶级数的展开级数，$t \in [0, T]$ 表示不同的采样位置。 $\left(L_x, L_y\right)$ 表示轮廓的中心点坐标，(a_n, b_n) 是傅里叶编码的x坐标，(c_n, d_n) 是傅里叶编码的 y 坐标。我们可以通过 FO 来粗略估计目标边界的复杂性。具体而言，当从解码的傅里叶级数得到的轮廓与原始轮廓达到一定程度的重叠时（使用 DICE 表示重叠的程度），此时解码的阶数定义为所需的累积次数。然而，当将此方法用作定量衡量指标时，设置适当的 DICE 阈值十分重要。DICE 设定阈值过低时不能准确区分各个目标边界之间的复杂性差异。如果阈值太高，EFD 可能为了拟合复杂轮廓而进行无限计算。因此，我们优化了 FO 的表示，以避免 EFD 计算过多的累计次数（参见公式 2 和公式 3 中的累积项）。

对于不同的解剖结构，我们设定 FO 从 1 开始增加，每增加 1 就计算一次解码轮廓与原始轮廓之间的 DICE。我们设定两种 EFD 结束的方式：1）DICE > 97.0%；2）DICE 在$F_{(a-1)}$系数和$F_{(a)}$系数之间的差异小于 0.1%。因此，我们在终止后记录 FO（$F_{(a)}$）和 DICE。最后，我们取 $F_{final} = F_a + n \times 100 \times (1 - DICE)$，$n = 2$作为最终的最优 FO。

本研究使用 Spearman 秩相关系数（Spearman's rank partial correlation coefficient）对上述目标结构的五个属性与 DICE 分数之间进行了偏相关性分析，同时，考虑了不同的测试策略。统计结果显示在表 6 中，而图 13 则展示了S_5策略的散点图。结果表示，在大多数测试策略中，DICE 分数在 FO 以及灰度差异呈现中等相关性（0.4 ≤ ρ < 0.7），与尺寸呈弱相关性（0.2 ≤ ρ < 0.4），与模态和长宽比无明显相关性。因此，在各种策略下，SAM 均可以稳定地分割具有不同模态和纵横比的医学解剖结构。同时，在框提示下，解剖结构大小可能会影响 SAM 的性能。此外，在处理具有复杂边界或低对比度特征的目标时，SAM 在所有测试策略下的性能往往较差。为了验证这个发现，我们在S_{5B}的策略下，将 DICE 平均分成了十个等级（例如，等级 1 表示 DICE（%）属于（0,10]），并在图 14 中可视化了不同 DICE 水平的 FO 箱线图。图中显示随着 DICE 水平的提高，结构的 FO 分布逐渐向较小值的范围偏移（证明 FO 越小的结构，DICE 值有越高的趋势）。此外，在图 15 中，我们展示了具有不同 FO 范围的解剖结构的可视化效果。这些可视化结果显示解剖结构的 DICE 分数在 FO 增加时倾向于降低，这也进一步表明了形状和边界复杂性可能会影响 SAM 的分割性能。

表 6. Spearman 偏相关分析（p < 0.001 的值以粗体显示）。

Strategy	Size		Intensity Difference		Fourier Order		Modality		Aspect Ratio	
	ViT-B	ViT-H	ViT-B	ViT-H	ViT-B	ViT-H	ViT-B	ViT-H	ViT-B	ViT-H
S_1	0.218	0.271	**0.503**	**0.614**	**-0.419**	**-0.453**	-0.006	0.048	0.056	0.035
S_2	0.236	0.293	**0.572**	**0.535**	**-0.537**	**-0.519**	0.051	0.049	0.079	0.072
S_3	0.289	0.330	**0.638**	**0.591**	**-0.520**	**-0.537**	0.094	0.065	0.046	0.053
S_4	0.275	0.339	**0.628**	**0.533**	**-0.524**	**-0.533**	0.062	0.040	0.048	0.060
S_5	**0.410**	**0.428**	**0.445**	**0.407**	**-0.479**	**-0.463**	-0.023	-0.034	0.065	0.076
S_6	0.370	0.392	**0.467**	**0.412**	**-0.621**	**-0.576**	0.014	-0.011	0.068	0.071

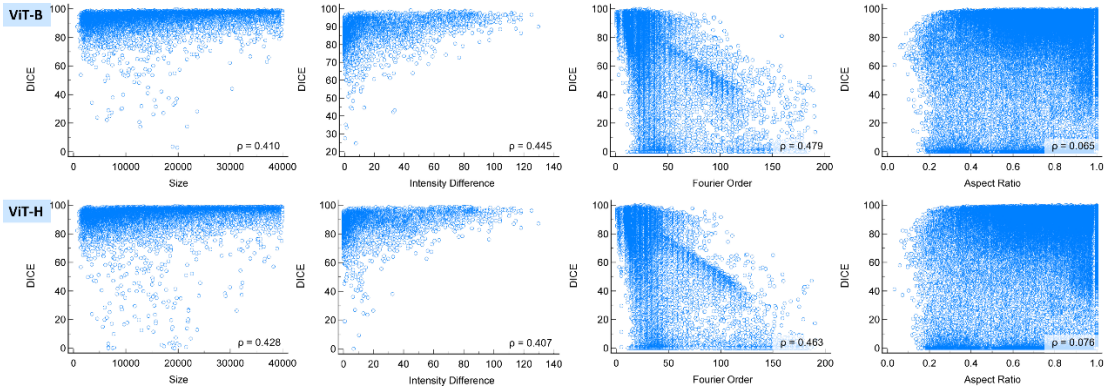


图 13. 不同目标属性在S_5策略下的 DICE 的散点图。

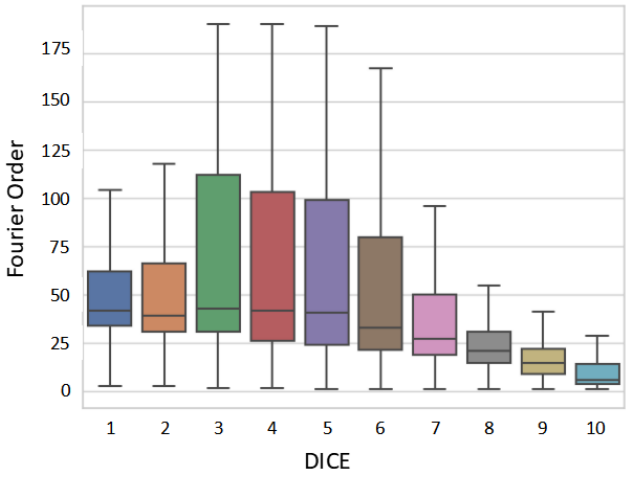


图 14. 不同 DICE 范围下的 FO 箱线图。

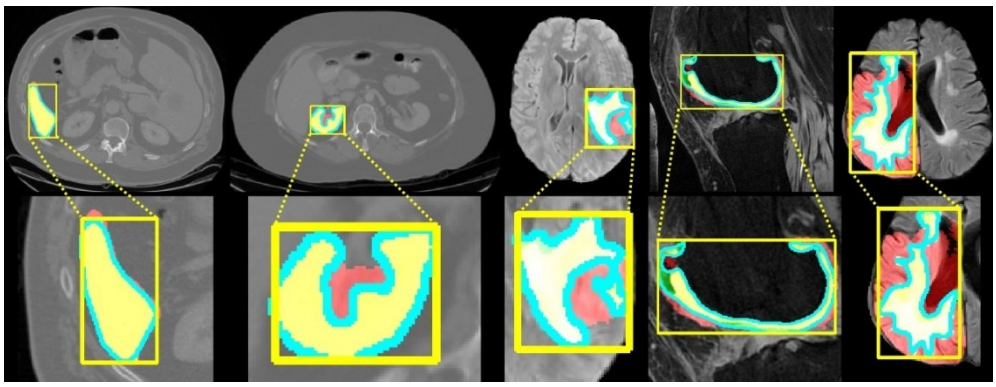


图 15. DICE 和 FO 之间的关系。从左到右，FO 逐渐增加。黄色框表示框提示，红色掩码是预测结果，绿色掩码是 GT，黄色掩码是预测和 GT 的重叠，蓝色轮廓是从傅里叶级数解码的。

4.9. 标注和质量分析

在本节中，我们讨论 SAM 是否能帮助医生减少标注时间并提升标注质量。我们在 COSMOS 1050K 中随机抽取了 100 张图像，SAM 对其预测结果的分割 DICE 在平均水平，以此构建了一个评估子集，包括 9 种不同模态的 55 个结构和 620 个掩膜，其中包括不同模态下相同目标。然后，我们邀请了三名具有 10 年经验的医生评估 SAM 的预测是否能提高标注速度和质量。他们的任务包括 1）从头开始标注评估子集中的所有目标，2）根据 SAM 的预测结果进行目标轮廓调整，3）记录两项任务的时间。为了评估注释质量，我们使用人工修正代价（Human Correction Efforts，HCE）指标[43]，该指标估计了将不准确的预测修正为满足实际应用中的准确结果（即 GT 掩膜）所需的人工标注工作量。HCE 指数越低表示掩膜（有/无 SAM 的人工注释）越接近 GT，即注释质量越高。如表 7 所示，通过 SAM 的帮助，可以实现更高的注释质量（HCE：0.27↓），并将注释速度提高约 25%。具体而言，标注一张图像平均可以节省约 1.31 分钟的时间，而标注一个目标结构，平均可以节省约 0.2 分钟的时间（因为上述任务中一张图像包含约 6.2 个目标）。值得注意的是，需要标记的解剖结构数量越多，SAM 效率的优势就越明显。

表 7. 医生在有/无 SAM 帮助的情况下的标注速度（s：秒，m：分钟）和标注质量。

SAM		Doctor	Human		Human with SAM	
HCE↓	Time (s)		HCE↓	Time (m)	HCE↓	Time (m)
5.66	0.47	Doctor1	4.74	4.41	**4.57**	**3.03**
		Doctor2	5.82	4.21	**5.23**	**2.95**
		Doctor3	4.65	4.19	**4.59**	**2.91**
		Mean	5.07	4.27	**4.80**	**2.96**

4.10. 不同提示随机性对性能的影响

在先前的实验中，我们固定了框和点的选择策略以保证实验的可重复性。我们通过选择质心和最小外接正框来测试 SAM 的理论最佳性能，因为它们可能包含关于目标的最具代表性的特征。然而，在真实场景中评估 SAM 时，准确点击每个目标的质心或绘制最小外接正框并不实际。因此，我们对于质心和框引入了不同程度的随机性，以模拟现实中的人的操作[44]。此外，我们认为这有助于更好地讨论 SAM 的稳定性。

具体而言，我们随机将框/点在 0-10、10-20 和 20-30 像素范围内进行放大/移动。在表 8 中，随机实验（Random 1-3）进行了三次，并计算了平均结果（Mean）。DICE 下降表示与没有偏移的原始结果相比，DICE 值的平均下降情况。对于S_2（单点），随着偏移水平的增加，DICE 性能下降了 2.67%、7.38%和 14.62%。随着点提示数量的增加（S_3和S_4），DICE 的下降可以得到缓解，并且模型的稳定性得到了提高。SAM 受框的偏移的影响很大（S_5，在 20-30 像素的偏移下性能下降了 24.11%），而在给框添加一个点的情况下，DICE 下降会进一步加剧（S_6，下降了 29.93%）。

表 8. 在不同的偏移级别和测试策略下的 DICE 下降程度的比较。

	Shift	DICE drop				
		S_2	S_3	S_4	S_5	S_6
Random 1	0-10	2.74	0.87	0.79	3.08	4.57
	10-20	7.55	1.42	1.37	10.24	13.98
	20-30	14.36	4.67	3.29	23.88	29.72
Random 2	0-10	2.68	0.90	0.84	3.24	4.49
	10-20	7.42	1.38	1.29	10.39	13.83
	20-30	14.51	4.49	3.17	23.71	29.50
Random 3	0-10	2.60	0.81	0.73	3.43	5.02
	10-20	7.17	1.19	1.22	10.87	14.28
	20-30	14.98	4.15	3.05	24.73	30.58
Mean	0-10	2.67	0.86	0.79	3.25	4.69
	10-20	7.38	1.33	1.29	10.50	14.03
	20-30	14.62	4.44	3.17	24.11	29.93

4.11. SAM 与传统交互式方法的比较

在前面的部分中，为了公平比较 SAM 的单轮推理性能，我们一次性将所有提示输入 SAM 的提示编码器。为了模拟真实的交互式分割过程，我们实现了基于多轮提示输入的 SAM。点的选择策略类似于常见的交互方法。具体而言，SAM 首先点击目标的质心，然后其余点的选择则基于 GT 和预测的假阴性（FN）和假阳性（FP）区域。然后，我们将 SAM 与两种不同的强交互分割方法进行了比较，即 FocalClick[45]和 SimpleClick[46]。它们都是在与 SAM 数据量相当的图像上进行预训练的。

我们选择了 10 个典型的器官/肿瘤，涵盖了各种模态、形状、大小和灰度分布。实验结果显示在图 16 中。基于 DICE 结果，我们的结论是：1）SAM 在第一轮单点交互中优于 FocalClick 和 SimpleClick；2）随着迭代的进行，SAM 的性能增长缓慢，甚至下降，而交互式方法的性能可以稳步提高；3）使用 10 个点时，SAM 的性能不如交互式方法。我们的结论与最近发表的 MedIA 论文[14]中一定的有共同之处。我们认为当前 SAM 在医学图像上基于点的多轮迭代能力相对较弱。未来的工作应该优化 SAM 的迭代训练策略，或者通过微调等方式来增强其多轮迭代的能力[47]。

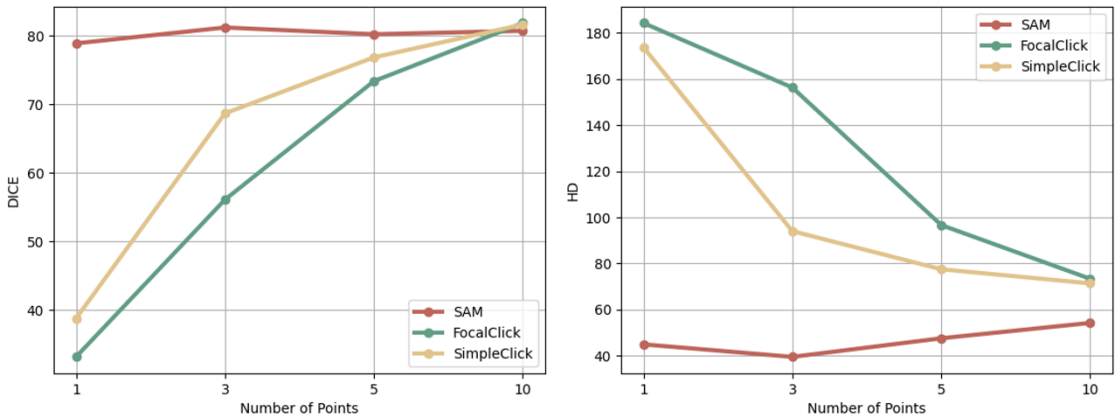


图 16. 三种不同方法随着点提示数量增加的平均性能变化。

4.12. 面向任务的 SAM 优化

SAM 在大多数医学图像/任务上的感知能力弱主要是由于医学相关训练数据的不足。SAM 的训练数据集，即 SA-1B[5]，包含了 1100 万张照片，包括自然场景、目标和场景等等，但不包括医学图像。自然图像通常与医学图像不同，前者具有颜色编码，目标的定义和边界相对清晰，更容易区分前景（目标）和背景（非目标），而且目标大小相对平衡。然而，大多数医学图像是灰度的，具有不清晰且复杂的结构边界，背景和前景相似，并且图像尺寸范围广泛（尤其包含一些非常小的目标）。

因此，我们使用 COSMOS 1050K 的部分数据对 SAM 进行微调，以提高 SAM 对医学目标的感知能力。具体而言，我们选择了 45 个常见和典型的目标，用于微调 SAM。受到 Ma 和 Wang 研究的启发[15]，我们只使用框提示对 SAM 进行微调。我们冻结了图像编码器以最小化计算成本。同时，提示编码器在编码框位置信息方面能力较为强大，所以，其参数也被冻结。因此，在微调期间，我们仅调整了掩膜解码器中的参数。我们将总的训练轮数设置为 20，学习率和批大小为 1e-4 和 2。

结果显示，在对 ViT-B 和 ViT-H 模型进行微调后，分割性能普遍得到了改善，如图 17 和图 18 所示。图 17 显示了在不同相关因素下观察，微调前后的 DICE 值都变得更高了（蓝色圈圈），表明整体性能有所提升。具体而言，如图 18 所示，对于 ViT-B，45 个目标中的 32 个显示出性能提升，而 ViT-H 在 45 个目标中有 37 个的性能有所提高。这证明了 ViT-H 的强大学习能力，因为它的参数几乎是 ViT-B 的 7 倍（636M / 91M）。此外，我们还发现对于具有数目较少、RGB 颜色编码等特征的目标性能会发生下降。这提醒我们可能需要针对特定数据集和任务对微调策略进行更仔细的设计。

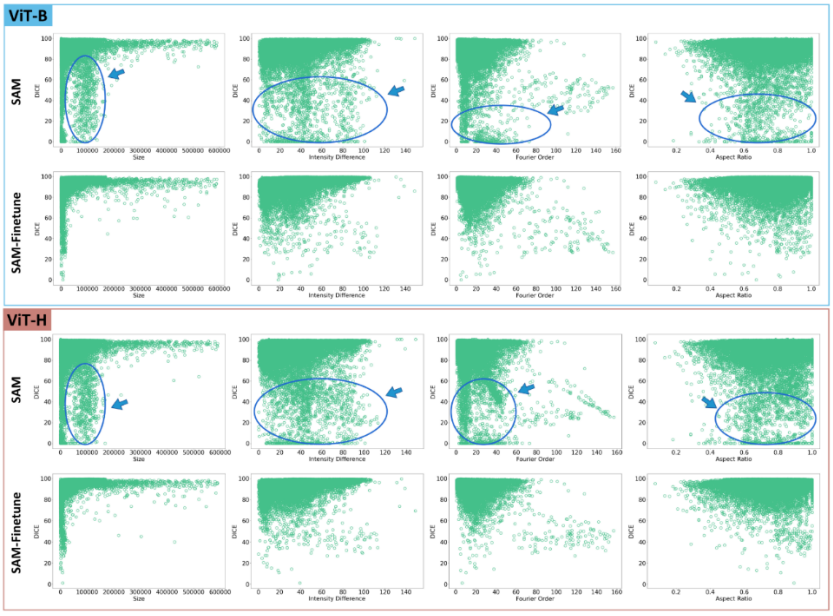


图 17. 在不同目标属性下 DICE 的趋势分析（ViT-B 和 ViT-H 模型，使用框提示的S_5测试策略）。蓝色圆圈表明了变化最明显的区域。

[5] https://ai.meta.com/datasets/segment-anything/

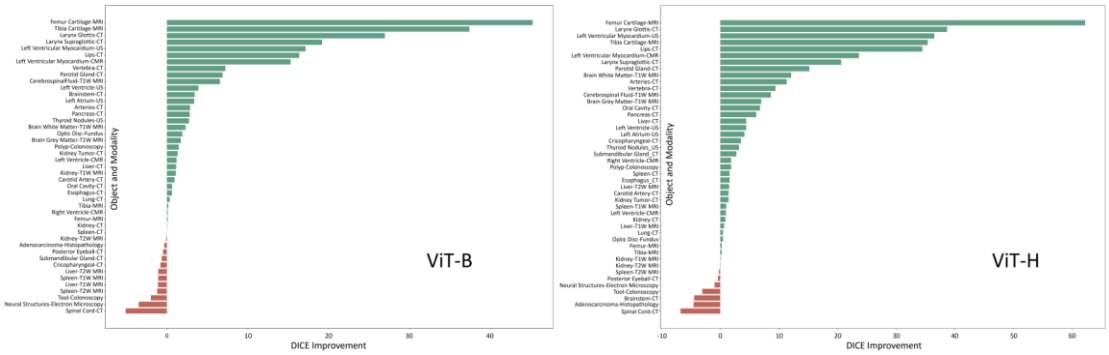


图 18. 微调前后的 DICE 的变化，左右分别为基于 ViT-B 和 ViT-H 微调的结果。

5. 结论

在本研究中，我们在大规模医学图像数据集上评估且分析了 SAM 分割性能。根据上述实验分析，我们的结论如下：1）SAM 在某些特定医学分割目标上表现出色，但在其他情况下表现不稳定、不完美甚至会完全分割失败。2）对比网络参数量较小的 ViT-B（91M）,基于参数量较大的 ViT-H（636M）的 SAM 在医学影像分割任务的整体性能上能得到提升。3）与 Everything 模式相比，SAM 在使用手动提示，特别是框提示时表现更好。4）SAM 可以较好地辅助人工专家标注，实现更高的标注质量和更少的标注时间。5）SAM 对点提示和框提示的随机偏移比较敏感，像素抖动程度的提升会造成越来越严重的分割性能下降。6）在少量点交互的情况下，SAM 比传统交互式方法表现得更好，但随着点提示数量的增加，SAM 的分割性能将会被超过。7）SAM 的性能与边界复杂性、灰度差异等因素相关- -边界越复杂、前景背景的差异越低，SAM 对医学目标的感知能力越差。8）使用医学数据对 SAM 进行微调，可使其平均 DICE 性能提高 4.39%（ViT-B）和 6.68%（ViT-H）。最后，我们相信，尽管 SAM 有望成为通用的医学图像分割模型，但其在医学图像分割任务中的性能目前还不够稳定。我们希望这份报告能帮助读者和社区更好地理解 SAM 在医学图像分割中的性能，并最终促进新一代医学影像分割基础模型的发展。

6. 讨论

我们将专注于讨论 SAM 潜在的未来发展方向，希望这些讨论能在一定程度上激发读者的思考。

在没有 GT 的情况下，SAM 如何获取语义信息？目前的 SAM 只具备感知物体的能力，无法分析物体的类别。最近的一些研究致力于解决这个问题，其中之一则在 SAM 中引入了 CLIP 模型[6]。具体而言，SAM 首先提供候选区域，然后从原始图像中裁剪出区域块。接下来，裁剪的区域块将输入到 CLIP 中进行目标分类。另一种解决方案是将 SAM 与开放词汇目标检测（Open-Vocabulary Object Detection，OVOD）模型结合，例如，通过将 DINO 与 SAM 进行

[6] https://github.com/Curt-Park/segment-anything-with-clip

组合（Grounded-SAM7[7]）。在这个框架中，OVOD 模型可以检测具有分类结果的目标边界框。然后，SAM 将采用框的区域作为输入并输出分割结果。最近，提出了语义 SAM[48]，用于在自然图像中分割和识别任何物体。所有先前的探索都是基于自然图像的。因此，开发具有语义认知的医学 SAM 是一个值得探索的领域。然而，这是具有挑战性的，因为在开放场景中的医学目标具有各种各样的形状，广泛的类型，以及许多相似的亚类（不同等级的肿瘤等）。

SAM 与传统分割方法相比？通过使用有限的医学数据对 SAM 进行微调，已经在性能上超越了一些面向特定任务的传统分割方法。这已在一些最近发表的研究中得到验证。在医学图像分割领域，对 2D SAM 进行微调在大多数情况下都能取得比经典 Unet 分割模型更为优越的性能[15]。同时 3D MA-SAM 已验证了通过使用 3D 适配器对 SAM 进行微调，在没有任何提示的情况下胜过传统的 SOTA 3D nn-Unet[49]。这为医学图像分割社区提供了启示--也许对基础分割模型进行微调会比从头开始训练传统分割模型表现更好。然而，SAM 仍然存在一些问题，包括模型对不同提示噪声/随机性的稳定性和多轮交互能力。

2D 还是 3D SAM? 对于医学数据，成像方式（2D/视频/3D/4D）的可变性可能会使一般模型的设计变得复杂。相对于视频/3D/4D 图像（CT/MRI 等），2D 在医疗数据中更为基础和常见。因此，构建一个能够一致处理所有类型数据的 2D 模型更为实用，因为视频/3D/4D 数据本质上都可以转换成一系列 2D 切片[15]。此外，有限的 3D 数据量（SAM: 11M 图像和 1B 掩模，而我们的:<10K 容积和<45K 掩模）可能会限制 3D 基础分割模型的构建，特别是如果需要从头开始训练一个 3D 模型。为了突破数据的限制，我们认为探索如何合成更多高保真度的三维数据是可行的，这有利于构建强大的医学三维影像分割基础模型。

SAM 如何助力大规模医学影像标注？大规模的完全标记的医学数据集对于发展强大的基于深度学习的医学分割模型至关重要。然而，目前的专家手动注释方案面临着极大的挑战。如在 Qu 等人的研究中[50]，他们对包含 9 个解剖结构和 320 万切片的 8,448 个 CT 体积进行标注，并评估得出结论：一个有经验的专家大致需要 30.8 年的时间。在这项工作中，作者通过多个预训练分割模型生成伪标签等策略，将注释时间缩短到三周。然而，获得性能良好的预训练模型，尤其是低假阳的模型仍然非常困难。此外，基于传统深度学习的分割网络不能很好地支持人机交互，限制了其灵活性。有了支持提示性分割的 SAM，带来了解决这些挑战的希望。我们的研究还初步验证了 SAM 可以显著缩短注释时间并提高注释质量。需要标记的解剖结构越多，SAM 效率的优势就越明显。值得注意的是，SAM 的设计范式具有实现通用分割的潜力。这意味着一个单一的 SAM 网络可以用于实现大规模多模态、多类别医学数据集的标注，而不是使用多个专门的任务模型。这对于模型在标注软件中的轻量和高效部署是至关重要的，例如 MONAI[51] 和 Pair 标注软件包[52]。

[7] https://github.com/IDEA-Research/ Grounded-Segment-Anything

7. 致谢

真诚感谢所有公开数据集的组织者和所有者的开源贡献，以及 Meta AI 公开发布了 SAM 的源代码。本研究的模型与代码已经开源：https://github.com/yuhoo0302/Segment-Anything-Model-for-Medical-Images，欢迎大家使用交流。

参考文献

[1] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[2] C. Jia *et al.*, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 4904–4916.

[3] A. Ramesh *et al.*, “Zero-shot text-to-image generation,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 8821–8831.

[4] A. Kirillov *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,* pp. 4015-4026.

[5] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *International Conference on Learning Representations*, 2020.

[6] L. Tang, H. Xiao, and B. Li, “Can SAM Segment Anything? When SAM Meets Camouflaged Object Detection,” *ArXiv Prepr. ArXiv230404709*, 2023.

[7] G.-P. Ji, D.-P. Fan, P. Xu, M.-M. Cheng, B. Zhou, and L. Van Gool, “SAM Struggles in Concealed Scenes–Empirical Study on" Segment Anything"," in *SCIENCE CHINA Information Sciences (SCIS)* 66 (12), 226101.

[8] W. Ji, J. Li, Q. Bi, W. Li, and L. Cheng, “Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-world Applications,” *ArXiv Prepr. ArXiv230405750*, 2023.

[9] S. Mohapatra, A. Gosai, and G. Schlaug, “Brain Extraction comparing Segment Anything Model (SAM) and FSL Brain Extraction Tool,” *ArXiv Prepr. ArXiv230404738*, 2023.

[10] R. Deng, C. Cui, Q. Liu, et al. Segment Anything Model (SAM) for Digital Pathology: Assess Zero-shot Segmentation on Whole Slide Imaging[C]//Medical Imaging with Deep Learning, short paper track. 2023.

[11] T. Zhou, Y. Zhang, Y. Zhou, Y. Wu, and C. Gong, “Can SAM Segment Polyps?,” *ArXiv Prepr. ArXiv230407583*, 2023.

[12] Y. Liu, J. Zhang, Z. She, A. Kheradmand, and M. Armand, “SAMM (Segment Any Medical Model): A 3D Slicer Integration to SAM,” *ArXiv Prepr. ArXiv230405622*, 2023.

[13] S. He, R. Bao, J. Li, P. E. Grant, and Y. Ou, “Accuracy of Segment-Anything Model (SAM) in medical image segmentation tasks,” *ArXiv Prepr. ArXiv230409324*, 2023.

[14] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, “Segment anything model for medical image analysis: an experimental study,” *Med. Image Anal.*, vol. 89, p. 102918, 2023.

[15] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, “Segment Anything in Medical Images.” arXiv, Jul. 17, 2023. doi: 10.48550/arXiv.2304.12306.

[16] J. Wu *et al.*, “Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation.” 2023.

[17] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[18] E. K. Wang, C.-M. Chen, M. M. Hassan, and A. Almogren, "A deep learning based medical image segmentation technique in Internet-of-Medical-Things domain," *Future Gener. Comput. Syst.*, vol. 108, pp. 135–144, 2020.

[19] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "UniverSeg: Universal Medical Image Segmentation," *in International Conference on Computer Vision*, 2023.

[20] A. L. Simpson *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *ArXiv Prepr. ArXiv190209063*, 2019.

[21] J. Shapey *et al.*, "An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI," *J. Neurosurg.*, vol. 134, no. 1, pp. 171–179, 2019.

[22] S. Bakas, M. Reyes, A. Jakab, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge[J]. 2018.

[23] Y. Sun *et al.*, "Multi-site infant brain segmentation algorithms: the iSeg-2019 challenge," *IEEE Trans. Med. Imaging*, vol. 40, no. 5, pp. 1363–1376, 2021.

[24] G. Podobnik, P. Strojan, P. Peterlin, B. Ibragimov, and T. Vrtovec, "HaN-Seg: The head and neck organ-at-risk CT & MR segmentation dataset," *Med. Phys.*, 2023.

[25] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?," *IEEE Trans. Med. Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[26] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant. Imaging Med. Surg.*, vol. 4, no. 6, p. 475, 2014.

[27] O. Viniavskyi, M. Dobko, and O. Dobosevych, "Weakly-supervised segmentation for disease localization in chest x-ray images," in *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*, Springer, 2020, pp. 249–259.

[28] A. A. A. Setio *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge," *Med. Image Anal.*, vol. 42, pp. 1–13, 2017.

[29] Z. Zhao, H. Chen, and L. Wang, "A coarse-to-fine framework for the 2021 kidney and kidney tumor segmentation challenge," in *Kidney and Kidney Tumor Segmentation: MICCAI 2021 Challenge, KiTS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings*, Springer, 2022, pp. 53–58.

[30] J. Ma *et al.*, "AbdomenCT-1K: Is Abdominal Organ Segmentation a Solved Problem?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6695–6714, 2022, doi: 10.1109/TPAMI.2021.3100536.

[31] Y. Ji, H. Bai, C. Ge, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 36722-36732.

[32] X. Luo *et al.*, "WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image," *Med. Image Anal.*, vol. 82, p. 102642, 2022.

[33] A. Sekuboyina *et al.*, "VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images," *Med. Image Anal.*, vol. 73, p. 102166, 2021.

[34] M. T. Löffler *et al.*, "A vertebral segmentation dataset with fracture grading," *Radiol. Artif. Intell.*, vol. 2, no. 4, p. e190138, 2020.

[35] S. Pang *et al.*, "SpineParseNet: spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation," *IEEE Trans. Med. Imaging*, vol. 40, no. 1, pp. 262–273, 2020.

[36] G. Lee, S. Kim, J. Kim, and S.-Y. Yun, "MEDIAR: Harmony of Data-Centric and Model-Centric for Multi-Modality Microscopy." 2022.

[37] J. Ma *et al.*, "The Multi-modality Cell Segmentation Challenge: Towards Universal Solutions," *in NeurIPS22 Cell Segmentation Challenge*, 2023.

[38] D. Jha *et al.*, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*, Springer, 2020, pp. 451–462.

[39] S. A. Hicks, D. Jha, V. Thambawita, P. Halvorsen, H. L. Hammer, and M. A. Riegler, "The EndoTect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*, Springer, 2021, pp. 263–274.

[40] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16000–16009.

[41] M. Tancik *et al.*, "Fourier features let networks learn high frequency functions in low dimensional domains," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 7537–7547, 2020.

[42] C. Mattjie *et al.*, "Zero-shot performance of the Segment Anything Model (SAM) in 2D medical imaging: A comprehensive evaluation and practical guidelines." 2023.

[43] X. Qin, H. Dai, X. Hu, D.-P. Fan, L. Shao, and L. Van Gool, "Highly accurate dichotomous image segmentation," in *European Conference on Computer Vision*, Springer, 2022, pp. 38–56.

[44] Y. Huang *et al.*, "Flip Learning: Erase to Segment," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, Springer, 2021, pp. 493–502.

[45] X. Chen, Z. Zhao, Y. Zhang, M. Duan, D. Qi, and H. Zhao, "Focalclick: Towards practical interactive image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1300–1309.

[46] Q. Liu, et al. Simpleclick: Interactive image segmentation with simple vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 22290-22300.

[47] J. Cheng *et al.*, "SAM-Med2D," *ArXiv Prepr. ArXiv230816184*, 2023.

[48] F. Li *et al.*, "Semantic-sam: Segment and recognize anything at any granularity," *ArXiv Prepr. ArXiv230704767*, 2023.

[49] C. Chen *et al.*, "MA-SAM: Modality-agnostic SAM Adaptation for 3D Medical Image Segmentation." 2023.

[50] C. Qu *et al.*, "Annotating 8,000 Abdominal CT Volumes for Multi-Organ Segmentation in

Three Weeks," *in Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[51] M. J. Cardoso *et al.*, "Monai: An open-source framework for deep learning in healthcare," *ArXiv Prepr. ArXiv221102701*, 2022.

[52] J. Liang *et al.*, "Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis," *Med. Image Anal.*, vol. 79, p. 102461, 2022.