TextSAM-EUS: 通过文本提示学习让 SAM 在内镜超声中准确分割胰腺肿瘤

Pascal Spiegler[1†], Taha Koleilat[1†], Arash Harirpoush[1], Corey S. Miller[2,3,4], Hassan Rivaz[1], Marta Kersten-Oertel[1], Yiming Xiao[1]
[1]Concordia University, Montreal, Canada
[2]Jewish General Hospital, Montreal, Canada
[3]McGill University Faculty of Medicine, Montreal, Canada
[4]Lady Davis Institute for Medical Research, Montreal, Canada
pascal.spiegler@mail.concordia.ca
[†]Equal contribution

Abstract

胰腺癌预后不佳，依赖于内镜超声（EUS）进行靶向活检和放射治疗。然而，EUS 的斑点噪声、低对比度以及直观性差使得通过完全监督的深度学习（DL）模型进行胰腺肿瘤分割变得容易出错，并且依赖于大量专家策划的注释数据集。为了解决这些挑战，我们提出了 TextSAM-EUS ，一种新颖、轻量化且无需手动几何提示的 Segment Anything Model（SAM）的文本驱动自适应版本。我们的方法通过 BiomedCLIP 文本编码器结合基于 LoRA 的 SAM 架构自适应，利用文本提示学习（上下文优化），实现 EUS 中胰腺肿瘤的自动分割，仅微调了 0.86 % 的总参数。在公共的胰腺内镜超声数据库上，TextSAM-EUS 使用自动提示获得了 82.69 % 的 Dice 系数和 85.28 % 的标准化表面距离（NSD），在使用手动几何提示时则达到了 83.10 % 的 Dice 系数和 85.70 % 的 NSD，优于现有的最先进（SOTA）监督 DL 模型和基础模型（例如 SAM 及其变体）。作为在 SAM 基础上首次尝试在医学图像分割中结合提示学习的方法，TextSAM-EUS 提供了一种高效且稳健的自动 EUS 分割的实际选择。我们的代码将在论文被接受后公开。

1. 介绍

胰腺癌是癌症相关死亡的第六大原因，生存率约为 10%，这凸显了其侵袭性以及对改进诊断和治疗策略的迫切需求。内镜超声（EUS）是一个医学成像技术，涉及将装备了超声探头的细长柔性管插入消化道，以获取胰腺等内脏器官的图像。该程序通过允许进行细针穿刺以获取组织样本和递送靶向癌症疗法（例如，插入放射性种子）在胰腺癌的临床管理中发挥重要作用。在这样的 EUS 程序中，肿瘤边界的准确描绘可极大地影响临床决策，因为精确的探头定位和覆盖是至关重要的。然而，斑点噪声、低对比度和 EUS 图像的不直观外观给自动分割带来了挑战。完全监督的架构，如基于 UNet 的模型，在医学图像分割，包括肿瘤描绘方面显示出强大的性能，但往往在对比度较高的图像模式（如 MRI 和 CT）上表现更好，而在噪声较大的超声上表现较差。此外，这些模型需要大量来自许多患者的像素级注释才能表现良好，这限制了它们在标记数据有限的临床场景中的实用性，例如从 EUS 中分割胰腺肿瘤。相对而言，Segment Anything Model（SAM），一个在超过十亿张图像上训练的基础模型，支持通过手动放置的几何提示（例如，点或框）进行零样本分割。然而，与几何提示相比，文本提示可以提供一种更为便捷的方式来启动分割，而无需对组织边界进行明确判断，因为它可以编码丰富的类级信息，但相关领域的探索尚有限。此外，SAM 的图像编码器完全在自然图像上进行了预训练，这在应用于医学图像特别是超声图像时导致了显著的领域偏移。因此，需要有效的参数高效微调方法以实现这些基础模型在医学图像上的应用 [28] 。

为了解决这些空白，我们提出了 TextSAM-EUS ，这是一种轻量级的 SAM 适应版本（仅调节其 0.86 % 的参数），专为 EUS 中的胰腺肿瘤分割而设计。我们的主要贡献如下：首先，我们开发了一种新颖的框架，该框架集成了文本提示学习（即，背景优化），以在参数高效的低秩适配（LoRA）微调 SAM 模型中启动 EUS 胰腺肿瘤分割。其次，我们引入了一个迭代分割优化步骤，该步骤通过几何提示（边界框和点）补充文本提示，以提高分割质量。第三，我们在一个公开的胰腺癌 EUS 数据集上对 TextSAM-EUS 进行了基准测试，其性能优于基于 UNet 的模型和其他基础模型，包括基于 SAM 的方法，使用完全自动化的文本驱动推理，实现了 82.69 % 的 Dice 相似系数（DSC）和 85.28 % 的归一化表面距离（NSD）。

Segment Anything Model（SAM）是一个基础模型，设计用于可提示的图像分割，它集成了一个强大的图像编码器、多功能的提示编码器和轻量级的掩码解码器，以支持零样本泛化。这种架构在医学影像界引起

Table 1. 在 SAM ViT-B 变体和所提出的 TextSAM-EUS 之间可训练参数与分割性能（DSC）的比较。TextSAM-EUS 在分割精度和参数效率之间达到了最佳平衡，以最小的参数量（1.69M）实现了最高的 DSC（82.69 %），同时在推理时不需要专家的手动干预。

Model	Manual Prompts?	Params. (M) ↓	DSC (%) ↑
SAM (Point) [19]	✓	93.74	39.80
SAM (Box) [19]	✓	93.74	78.15
SAMUS [24]	✓	39.65	70.75
MedSAM [28]	✓	93.74	82.66
SAMed [36]	✗	1.45	78.00
AutoSAM [30]	✗	41.56	81.26
AutoSAMUS [24]	✗	8.86	81.04
TextSAM-EUS (Ours)	✗	1.69	82.69

了极大的兴趣。例如，MedSAM 对大约一百万对医学图像-掩码进行了微调，在多种分割任务上取得了优异的表现。为提高效率，AutoSAM 引入了一种替代的微调策略，仅训练提示编码器，并使用专为医学应用定制的基于反卷积的解码器。同样，Zhang 等人提出了 SAMed，它在 SAM 的图像编码器上附加了 LoRA 适配器，并微调提示/掩码解码器，从而在最少的可训练参数下实现全自动医学图像分割。此外，Wu 等人引入了 SPFS-SAM，为 SAM 配备了一个自提示机制，其中一个轻量级分类器从 SAM 的嵌入中生成初始掩码，然后自动生成点提示以进行迭代优化。此外，Cheng et al. [5] 系统地评估了不同的提示类型，确定边界框在 12 项任务的医学分割中最为有效。为解决 SAM 的嘈杂伪标签问题，Huang et al. [12] 提出了一个校正机制，可以优化这些标签以改善后续的微调。在另一个方向，Gong et al. [8] 通过用 3D 卷积模块替代 SAM 的掩码解码器调整模型以处理 3D 体积医学数据。此外，MedCLIP-SAM [21, 22] 通过将 SAM 的视觉特征与 MedCLIP 的医学文本嵌入对齐，将视觉-语言预训练引入 SAM，在不需要密集标注的情况下增强零样本分割。此外，最近针对颅内出血分割的具不确定性意识的 SAM 适配 [32]，将基于 YOLO 的检测与经过不确定性校正的 SAM 框架相结合，以应对弱监督场景。尽管这些适配扩展了 SAM 在医学领域的适用性，它们常常依赖于再训练程序、提示工程或手动几何提示。相比之下，我们的方法利用 BiomedCLIP 生成的文本提示，实现无需手动提示的推理，不需要点、框或掩码。

1.1. 超声胰腺肿瘤分割

之前的几项研究探索了在超声（US）图像中分割胰腺肿瘤的这一具有挑战性的任务。传统的监督方法，例如由 Lu 等人 [27] 和 Huang 等人 [11] 提出的，利用了改进的 U-Net 架构。Lu 等人引入了一种多尺度注意力 U-Net 以在嘈杂的超声图像中聚焦于肿瘤区域，而 Huang 等人则采用了一种带有形状约束的级联分割框架来改善边界描绘。半监督方法也被研究过，特别是由 Liu 等人 [25]，他们提出了一种多任务一致性学习策略，以利用标记和未标记数据，解决医学超声数据集标注稀缺的问题。最近，基础模型已被应用于超声分割。SAMUS [24] 通过额外的 CNN 编码器微调 SAM，在 US 图像上展示出良好的结果。AutoSAMUS [24] 基于 SAMUS 训练网络从图像编码器特征中自动生成其点提示，消除了手动输入的需要。CC-SAM [9] 通过使用 grounding DINO 获得几何提示文本进一步提高了 SAMUS 的泛化能力。这些努力强调了一种日益增长的趋势，即结合强大的基础模型与超声特定的适应来增强在低对比度、噪声成像模式（如超声）中的分割性能。

1.2. 生物医学领域的 VLMs

视觉-语言模型（VLMs）如 CLIP [29] 和 ALIGN [15] 通过使用对比自监督技术将图像和文本表示映射到一个联合嵌入空间，进一步推动了多模态学习的发展。虽然这些模型在一般领域中擅长于零样本分类和跨模态检索等任务，但在医学等专业领域中，其性能常常下降，因为这些领域重视细微的视觉特征和特定领域的语言。为了克服这些限制，近期的工作探索了领域适应策略，其中，提示学习作为模型完全微调的计算高效替代方案出现。方法如 CoOp [42] 和 CoCoOp [41] 通过学习特定任务的文本提示标记来优化文本上下文，同时保持基础 VLM 固定。进一步的发展，如 MaPLe [16] 和 PromptSRC [17]，通过引入编码器调整和正则化策略来增强泛化能力。基于适配器的方法，包括 CLIP-Adapter [7] 和 Tip-Adapter [37]，通过优化视觉编码器或利用支持集来改进小样本学习，尽管它们可能会遇到训练不稳定的问题。在生物医学领域，一些 CLIP 变体，如 BioViL [1]、PubMedCLIP [6] 和 BiomedCLIP [38]，已结合医学语料库以改善领域对齐。其他工作 [21, 22] 进一步将这些通用 VLM 与临床影像应用相连接。然而，捕捉细粒度临床语义仍然是一个重大挑战 [35, 40]。最近的努力已将 CoOp 风格的提示适配到医学领域；DCPL [4] 就是一个显著的例子，尽管这些通常需要中到大型的训练集。相对而言，BiomedCoOp [23] 表明，通过文本提示调优，同时保持任务的泛化性，仍然可以在低资源医学场景中实现出色的性能。尽管在单独适配 CLIP 和 SAM 用于医学任务方面已有进展，但在生物医学领域内结合使用这两个模型，特别是高效提示学习的统一端到端架构仍未被探索。

2. 方法与材料

在图 1 中展示了所提出的 TextSAM-EUS 框架的总体概述，其中包括具有 BiomedCLIP 文本编码器的可学习上下文标记、SAM 图像编码器和掩码解码器的 LoRA 适应，以及一个迭代分割细化模块。

2.1. SAM 预备知识

Segmentation Anything 模型 [19] 包括三个核心组件：大型图像编码器、提示编码器和轻量级掩码解码器。
图像编码器：图像编码器 $\boldsymbol{E}_{\text{img}}$ 是一个基于 ViT 的主干网络，用于处理输入图像 $\boldsymbol{X} \in \mathbb{R}^{3\times H\times W}$ 以生成特征

图：

$$\boldsymbol{F} = \boldsymbol{E}_{\text{img}}(\boldsymbol{X}) \in \mathbb{R}^{h \times w \times d} \tag{1}$$

其中 $h \times w$ 是编码特征网格的分辨率，d 是特征维度。提示编码器：提示编码器 $\boldsymbol{E}_{\text{prompt}}$ 将用户定义的提示（例如点、边界框或掩码）映射为提示标记：

$$\boldsymbol{P} = \boldsymbol{E}_{\text{prompt}}(\text{ prompt }) \in \mathbb{R}^{k \times d} \tag{2}$$

其中 k 取决于所提供提示的数量和类型（例如，点的数量或框角）。
掩码解码器：解码器 $\mathcal{D}_{\text{mask}}$ 将图像特征 $\boldsymbol{F}$ 与提示嵌入 $\boldsymbol{P}$ 融合，以预测分割掩码：

$$\hat{\boldsymbol{Y}} = \mathcal{D}_{\text{mask}}(\boldsymbol{F}, \boldsymbol{P}) \in \mathbb{R}^{H \times W} \tag{3}$$

这种架构使得 SAM 能够基于稀疏或密集的几何提示生成分割掩码，支持跨不同领域的通用分割。

为了实现文本驱动的分割，我们利用了 CLIP [29] 架构（更具体地说是 BiomedCLIP [38] ），该架构由一个视觉编码器 $\boldsymbol{E_v}$ 和一个文本编码器 $\boldsymbol{E_t}$ 组成，将图像和文本映射到一个共享的嵌入空间中。在我们的框架中，我们使用预训练的 BiomedCLIP 文本编码器 $\boldsymbol{E_t}$ 来编码特定类别的文本提示，并将其输入到 SAM 的提示编码器中，有效地根据自然语言调整分割过程。

给定一批 B 图像和 C 文本类提示，视觉输入表示为 $\boldsymbol{X_v} \in \mathbb{R}^{B \times 3 \times H \times W}$ ，代表大小为 $H \times W$ 的 RGB 图像，文本输入表示为 $\boldsymbol{X_t} \in \mathbb{R}^{C \times L}$ ，其中 L 是标记化的序列长度。这些被处理为：

$$\boldsymbol{T} = \boldsymbol{E_t}(\boldsymbol{X_t}) \in \mathbb{R}^{C \times D} \tag{4}$$

，其中 $\boldsymbol{T}$ 包含用于引导 SAM 中的提示编码器的语言嵌入，且 D 是嵌入维度。

这种方法使得 SAM 可以根据文本进行条件设定，从而允许其基于自然语言提示对图像区域进行分割。通过将 BiomedCLIP 的语言先验纳入 SAM 提示编码器中，我们的方法使文本驱动的分割功能能够从更好的生物医学语义中受益。

尽管 SAM 在自然图像上表现出强大的泛化能力，但它并未在医学图像上进行预训练，因此无法捕捉超声扫描中存在的微妙解剖变化。尤其是在超声波中，胰腺肿瘤的外观变化很大且对比度低，阻碍了强健的零次分割。为了缓解这种领域差异，我们将 LoRA 模块应用于 SAM 的图像编码器和掩码解码器。LoRA 不是更新整个权重矩阵，而是为线性层引入可训练的低秩更新。对于一个权重矩阵，其适应版本变为：其中，表示适应的内在秩，A 和 B 包含可训练参数。该设计允许在保持原有模型大部分冻结的同时，SAM 对超声数据进行高效适应。LoRA 模块与接下来描述的语言引导提示调整一起进行优化。

为了使 SAM 能够执行基于生物医学文本的分割，我们提出了一种基于 CLIP 的文本提示学习策略。我们没有使用更常见的几何提示，而是在上下文优化 [18] 的基础上直接对自然语言查询进行编码，使用 BiomedCLIP [38] 编码器（PubMedBERT [10] ），然后将这一语义表示融入 SAM 的提示编码器中以启动分割。

在这里，我们引入了 b 个可学习的提示标记 $\{P^i \in \mathbb{R}^{d_t}\}_{i=1}^{b}$ ，其中 d_t 是 BiomedCLIP 文本编码器的隐藏维度。这些提示标记与 S 个固定标记（表示为 $W_0 = [w^1, w^2, \cdots, w^S]$ ）连接在一起，从而使标记总数保持为 N ，满足文本编码器的固定上下文长度（即 $S + b = N$ ）。结果序列通过 K 个顺序 Transformer 层传递：

$$[P_0, W_0] = [P^1, P^2, \cdots, P^b, w^1, w^2, \cdots, w^S], \tag{5}$$

$$[\_, W_i] = \mathcal{L}_i([P_{i-1}, W_{i-1}]), \quad i = 1, \ldots, t, \tag{6}$$

$$[P_j, W_j] = \mathcal{L}_j([P_{j-1}, W_{j-1}]), \quad j = t+1, \ldots, K \tag{7}$$

$$\boldsymbol{z} = \text{TextProj}(w_K^S), \tag{8}$$

其中 $\mathcal{L}_i$ 表示第 i 个 Transformer 块，TextProj$(\cdot)$ 将最终的 [EOS] 标记表示投影到与图像编码器共享的固定嵌入空间中。对于 BiomedCLIP，$N = 256$ ，并且最终标记 w_K^S 对应于 [EOS] 标记。提示标记只在前 t 层中被主动更新，这定义了文本提示的深度。所得的文本嵌入 $\boldsymbol{z} \in \mathbb{R}^D$ 能够捕捉到丰富的生物医学语义。

我们使用最终的文本嵌入 $\boldsymbol{T} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_C] \in \mathbb{R}^{C \times D}$ 作为 SAM 提示编码器的输入，而不是几何提示：

$$\boldsymbol{P}_{\text{text}} = \boldsymbol{E}_{\text{prompt}}^{\text{text}}(\boldsymbol{T}) \in \mathbb{R}^{C \times d}, \tag{9}$$

，其中 $\boldsymbol{E}_{\text{prompt}}^{\text{text}}$ 是一个轻量级适配器，作为带有 GELU 激活的两层 MLP 实现：

$$\boldsymbol{E}_{\text{prompt}}^{\text{text}}(\boldsymbol{z}_i) = \boldsymbol{W}_2 \cdot \text{GELU}(\boldsymbol{W}_1 \boldsymbol{z}_i + \boldsymbol{b}_1) + \boldsymbol{b}_2, \tag{10}$$

，有参数 $\boldsymbol{W}_1 \in \mathbb{R}^{m \times h}$ 、$\boldsymbol{W}_2 \in \mathbb{R}^{m \times m}$ 、$\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^m$ ，以及尺寸 $h = 512$ 和 $m = 256$ 。此转换确保了 BiomedCLIP 文本嵌入与 SAM 提示标记空间之间的兼容性。最后，这些由文本生成的提示标记与从图像编码器提取的预训练密集掩码嵌入连接，形成完整的 SAM 解码器输入提示集：

$$\boldsymbol{P} = [\boldsymbol{P}_{\text{text}}; \boldsymbol{P}_{\text{dense}}] \tag{11}$$

。此完整的提示标记集 $\boldsymbol{P}$ 随后传递到 SAM 的掩码解码器 $\mathcal{D}_{\text{mask}}$ ，后者在 $\boldsymbol{P}$ 和图像嵌入 $\boldsymbol{F}$ 之间执行交叉注意力以预测最终分割掩码。

为了提高在具有挑战性的区域的分割精度，我们在初始 SAM 掩码预测之后应用了一个轻量级的分割精细化步骤。我们提取几何线索，特别是预测区域的边界框和中心点，以生成几何提示（框和点）。这些几何提示与从文本中得出的提示标记和密集掩码嵌入相连接，形成一种混合提示方法，结合语义指导和几何特异性，使得 SAM 能够在最小的计算开销下精细化其预测。在我们的实验中，我们使用了两次分割精细化迭代：

$$\boldsymbol{P}_{\text{post}} = [\boldsymbol{P}_{\text{text}}; \boldsymbol{P}_{\text{dense}}; \boldsymbol{P}_{\text{geometric}}] \tag{12}$$

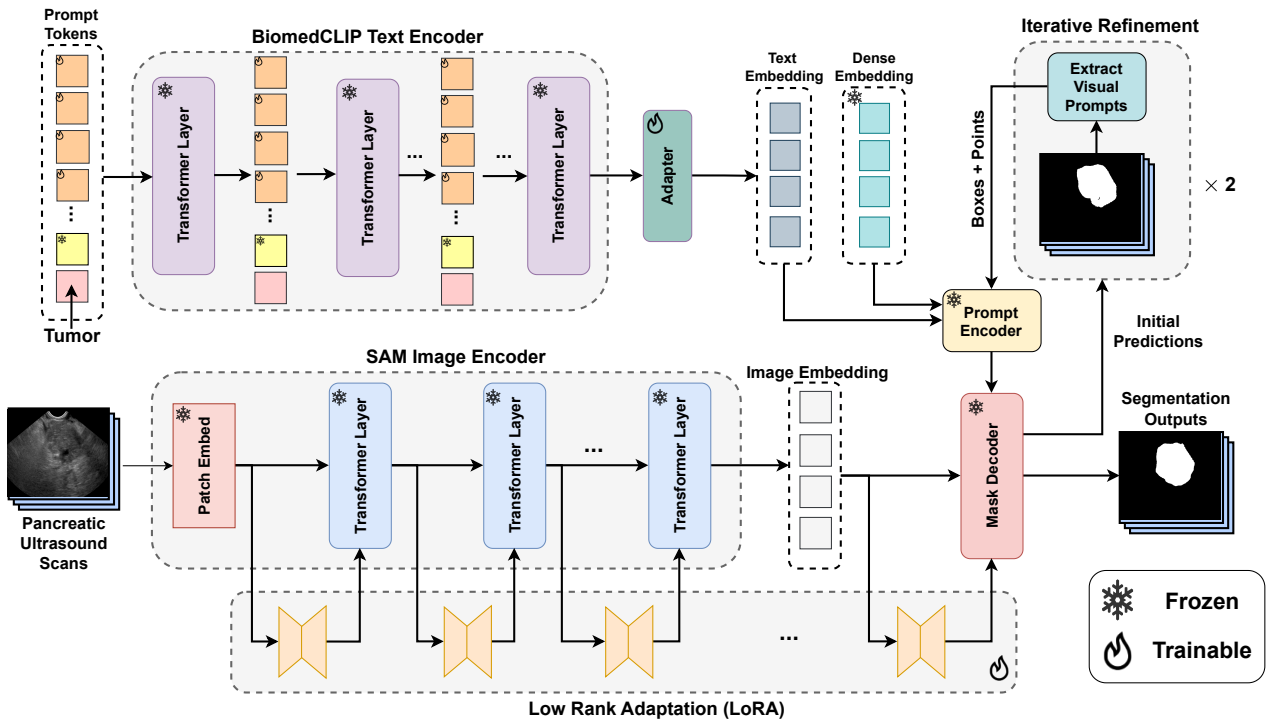

Figure 1. TextSAM-EUS 框架概述。我们对 BiomedCLIP 文本编码器 $\boldsymbol{E}_t$ 中的上下文标记进行微调，以生成文本嵌入 $\boldsymbol{T}$ ，这些嵌入通过适配器 $\boldsymbol{E}_{\text{prompt}}^{\text{text}}$ 投射到 SAM 提示标记空间。这些嵌入引导 SAM 掩码解码器 $\mathcal{D}_{\text{mask}}$ ，该解码器将它们与来自基于 ViT 的图像编码器 $\boldsymbol{E}_{\text{img}}$ 的图像特征 $\boldsymbol{F} = \boldsymbol{E}_{\text{img}}(\boldsymbol{X})$ 融合，图像编码器使用 LoRA 进行了适配。迭代分割优化步骤结合了初始预测中的几何提示（框和点），以提高分割质量。

3. 实验与结果

我们使用公开可用的胰腺内窥镜超声数据库 [14] 评估我们的方法，该数据库包含在 EUS 引导程序中获取的高分辨率超声图像。该数据集包括由临床专家标注的胰腺肿瘤区域，提供了用于分割任务的宝贵像素级真值。为了确保评估的鲁棒性并避免数据泄漏，我们对数据集中所有患有胰腺肿瘤的 18 名患者进行了严格的患者划分。最终的训练集包含来自 12 名患者的 11,363 张 EUS 图像。验证集包含来自另外 2 名患者的 986 张图像，而分割测试集则包括从 4 名未参与训练的肿瘤患者中收集的 4,185 张图像。所有图像均为灰度，分辨率为 711 × 457 像素。

3.1. 基线模型

我们将 TextSAM-EUS 与三类深度学习方法进行比较，这些方法均在相同的胰腺 EUS 数据集划分上进行训练和评估（见第 3 节）。首先，我们训练了两个流行的全监督模型，nnUNet 和 SwinUNet。nnUNet [13] 能够自动配置 U-Net 架构以及数据集的最佳性能所需的预处理和后处理，而 Swin-UNet [3] 则将一个移窗 Transformer 编码器集成到类似 U-Net 的配置中，以捕获全局上下文和局部细节。其次，我们评估了两个非 SAM 基础模型，包括 BiomedParse [39] ，一个在生物医学图像-文本对上预训练并适用于基于提示分割的视觉语言模型，以及 UniverSeg [2] 。最后，我们对比评估了 SAM 和五种基于 SAM 的方法：SAMed [36] 、AutoSAM [30] 、AutoSAMUS [24] 、SPFS-SAM [33] 和 MedSAM [28] ，其架构在第 ?? 节中描述。我们在训练集上微调了 SAMed、AutoSAM、AutoSAMUS 和 SPFS-SAM，而 MedSAM 和原始 SAM 则使用其预训练权重进行评估。对于 SAMUS，由于其是一个可调基础模型，我们报告了使用预训练和微调权重的结果。

对于期望点提示的基线模型（SAMUS，SAM），我们从每个真实值中随机采样一个点。对于期望边界框的模型（MedSAM，SAM），我们从真实值中提取紧密的边界框，然后每条边扰动 10-15 个像素，以模拟人为提供的输入。由于 TextSAM-EUS 的细化步骤使用了一个点和一个框，为了测试其在手动提示下的性能，我们为手动提示变体提供了一个真实值点和一个扰动后的边界框。对于 BiomedParse [39] ，在推理时需要文本提示，我们为每张图像提供固定的文本提示 “ pancreatic tumor in ultrasound scan “。对于 UniverSeg，我们进行了 16-shot 和 32-shot 微调实验。所有训练的基线模型均使用其原始论文推荐的超参数设置进行了优化。我们提出的 TextSAM-EUS 模型在训练集上训练了 5 个时期，使用学习率 0.001 和批量大小 1。我们使用

AdamW 优化器 [26]，权重衰减为 0.01，并采用余弦退火学习率计划。模型选择基于最低的验证损失。我们使用由 Dice 损失和二元交叉熵（BCE）损失组成的复合损失函数。BiomedCLIP 文本编码器配置了 4 个上下文标记和 12 层深度（贯穿提示编码器的前 12 个 Transformer 层）。我们使用 “ tumor ” 作为类名称，随机初始化上下文标记（参见图）。1 ），并应用 LoRA 进行参数高效微调，其秩为 16。我们使用 Dice 相似系数和归一化表面距离（在 3 像素的边界容差下计算）评估分割性能，结果以 0-100 % 范围内的平均值 ± 标准差报告，涵盖所有测试案例。使用双尾配对样本 t 检验比较方法之间的分割性能，其中 p 值 < 0.05 表示差异在统计学上有显著性。训练是在一台 NVIDIA A100 GPU（40GB RAM）上完成的。

3.2. 分割结果

表 2 展示了在保留的测试集上的 DSC 和 NSD。方法根据第 3.1 节定义的基线类别进行分组。TextSAM-EUS 在不需要手动提示的方法中实现了最高的 DSC (82.69 %) 和 NSD (85.28 %) ($p < 0.05$)。与最强的自动 SAM 变体（AutoSAM 和 AutoSAMUS）相比，我们的模型分别提高了 DSC $\sim$ 1.4 % 和 1.65 % ，同时调整的参数要少得多（1.69 M vs. 41.56 M 和 1.69 M vs. 8.86 M）。MedSAM 和自动 TextSAM-EUS 得到几乎相同的 DSC (82.69 vs. 82.66; $p = 0.906$)，而 MedSAM 在 NSD 上显示出轻微的优势 (85.28 vs. 85.75; $p = 0.035$)。然而，需要注意的是，MedSAM 在推理时配备了真实值边界框。还需要指出的是，我们的方法明显优于 SAMed [36] ，后者也利用 LoRA 来调整图像编码器和蒙版解码器模块，但仅调整密集嵌入，而我们的方法注入丰富的生物医学文本语义以提升分割性能，强调了深度文本提示的益处。当配备额外的手动几何提示时，TextSAM-EUS 获得了显著高于 MedSAM 的 DSC ($p = 0.039$), 尽管其在 NSD 方面的差异不显著 ($p = 0.801$), 尽管平均 NSD 略高。这些结果表明，我们的自动文本驱动 TextSAM-EUS 明显优于所有自动基线，并且在配备手动提示时与其他手动提示模型相匹配或超越。

3.3. 消融实验

为了确定 TextSAM-EUS 中每个组件的贡献，我们进行了四项消融研究，分别针对以下方面：(1) LoRA 适应性秩；(2) 可学习上下文标记的长度；(3) 处理这些标记的深度；以及 (4) 后处理阶段几何提示的选择。所有结果均报告在分割测试集上（平均值 ± 标准差）。

为了确定有效的肿瘤分割所需的适应能力，我们将 LoRA 秩从 1 变动到 16（见表 3 ）。在 $r = 1$ 时，模型无法学习到有意义的特征（44.00 % DSC），而在 $r = 4$ 时性能有所恢复（81.33 % DSC, 83.19 % NSD）。进一步扩大到 $r = 16$ 时产生了较小但稳定的改进（82.69 % DSC, 85.28 % NSD），这表明当适配器捕获到胰腺肿瘤的主要外观变化时，其能力趋于饱和。

Table 2. 模型之间的分割性能比较。DSC 和归一化表面距离（NSD）指标以均值 ± 标准差的形式报告。使用真实值提示的模型在推理时代表上限。

Model Name	DSC (%) ↑	NSD (%) ↑
UNet-based Approaches		
nnUNet	79.94 ± 19.96	82.30 ± 20.24
SwinUNet	59.69 ± 23.33	10.22 ± 6.64
Foundation Models		
BiomedParse	37.00 ± 37.76	38.92 ± 36.81
UniverSeg (16 shots)	42.92 ± 6.55	47.54 ± 6.86
UniverSeg (32 shots)	43.61 ± 6.70	48.03 ± 7.00
SAM-based Approaches (Automatic Prompts)		
SAMed	78.00 ± 16.95	80.62 ± 17.00
AutoSAM	81.26 ± 16.98	83.79 ± 17.19
AutoSAMUS	81.04 ± 16.49	83.80 ± 16.61
SPFS-SAM	60.64 ± 9.69	63.13 ± 9.82
TextSAM-EUS (Ours)	82.69 ± 15.14	85.28 ± 15.21
SAM-based Approaches with Point/Box Prompts (Upper Bounds)		
SAMUS	70.76 ± 16.60	73.86 ± 16.57
SAMUS (Finetuned)	82.81 ± 12.56	85.50 ± 12.72
MedSAM	82.66 ± 6.80	85.75 ± 6.28
SAM (Point)	39.80 ± 22.78	41.54 ± 23.83
SAM (Box)	78.15 ± 18.06	83.70 ± 18.06
TextSAM-EUS (Ours)	83.10 ± 14.13	85.70 ± 14.12

Table 3. 关于 LoRA 秩对分割性能影响的消融研究。一般来说，更高的秩会提高性能，其中秩 16 取得了最佳结果。

LoRA Rank (r)	Params. #	DSC (%) ↑	NSD (%) ↑
1	1,043,812	44.00 ± 28.98	46.99 ± 29.34
4	1,172,068	81.33 ± 16.53	83.19 ± 16.66
16	1,685,092	82.69 ± 15.14	85.28 ± 15.21

3.3.1. 上下文标记长度的影响

表格 4 检查了可学习的上下文提示的长度 b ，表明四个标记提供了最佳的 DSC 和 NSD 得分。增加到八个标记会引入优化不稳定性并降低准确性，而 12 个标记则部分恢复性能，但不超过四个标记的设置。这些观察表明，一个简洁的提示可以为目标二元任务编码足够的语义细节。

Table 4. 关于上下文标记数量对分割性能影响的消融研究。

Context Length (b)	DSC (%) ↑	NSD (%) ↑
4	82.69 ± 15.14	85.28 ± 15.21
8	66.74 ± 21.68	68.29 ± 21.86
12	82.10 ± 15.20	83.99 ± 15.27

3.3.2. 提示注入深度的效果

BiomedCLIP 文本编码器包含 12 个 Transformer 层。提示注入深度 t 指示了多少个前期层（层 1 到 12 ）接收可调整的深度提示令牌，剩余的层保持不变。表 5 显示，随着更深的处理，性能持续上升，并在提示贯穿所

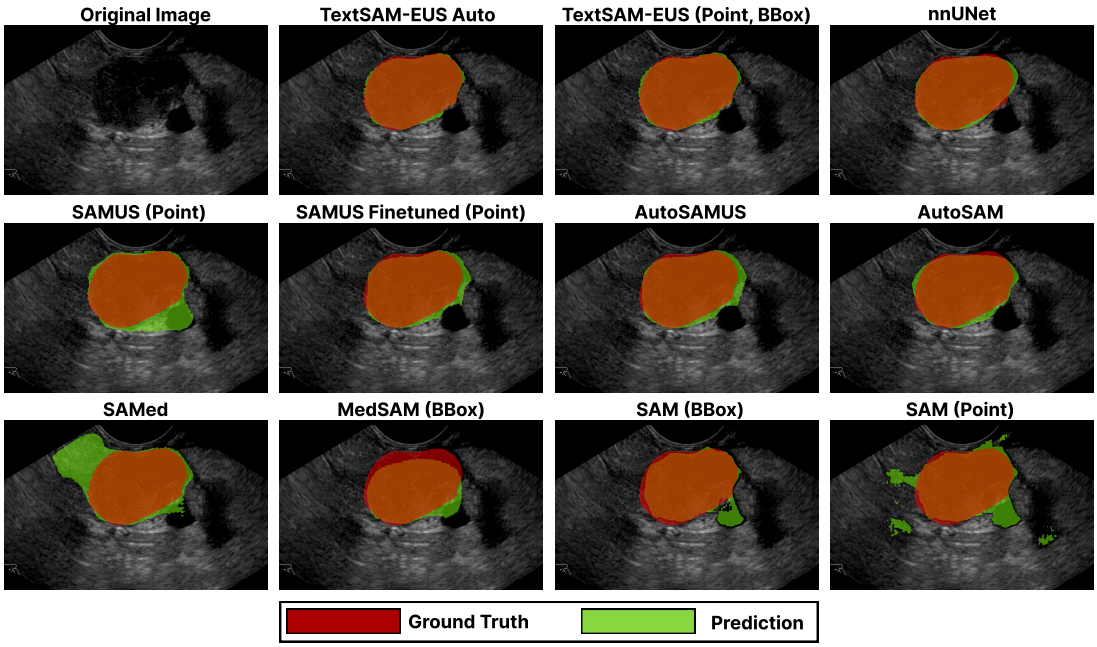

Figure 2. 在一个具有代表性的 EUS 切片上进行定性比较，比较九个最具竞争力的基线与 TextSAM-EUS 的全自动和人工提示变体。绿色表示预测的掩码，红色表示真实值，橙色表示二者重叠的区域。支持人工提示的模型在括号中描述提示类型，其余为自动的。

有层（$t = 12$）时达到峰值，这证实了上下文令牌必须影响大部分编码器才能有效引导分割。

Table 5. 关于可学习提示深度对分割性能影响的消融研究。

Layer Depth (t)	DSC (%) ↑	NSD (%) ↑
1	78.32 ± 19.05	80.97 ± 19.29
4	76.22 ± 19.47	78.60 ± 19.67
9	80.52 ± 17.71	82.36 ± 17.85
12	82.69 ± 15.14	85.28 ± 15.21

3.3.3. 几何提示在精炼阶段的效果

表 6 评估了从我们的文本提示预测掩码中自动提取的几何线索。具体而言，我们计算了由 TextSAM-EUS 输出的初始掩码的边界框和质心，并将这些几何衍生的提示反馈到细化阶段。第一行代表在没有迭代分割细化阶段的情况下框架的性能，即仅使用语言引导的预测掩码。即使没有几何指导，仅语言掩码也能达到 82.00% DSC。仅使用提取的边界框或质心之一可以取得适度的改进。当自动衍生的箱体和质心提示与学习的文本提示结合时，我们观察到了最高的 DSC 和 NSD 分数，添加额外的随机点没有进一步的益处。

3.3.4. 总结

在所有研究中，我们观察到适中的适配器容量（$r = 16$）、简洁的提示（$b = 4$）、深度提示注入（$t = 12$）以及单一重心点 + 边界框分割优化策略在参数效率和分割精度之间提供了最佳的权衡。

Table 6. 分割细化阶段中不同几何提示的效果

BBoxes	Points	DSC (%) ↑	NSD (%) ↑
✗	✗	82.00 ± 17.40	84.61 ± 17.53
✓	1 Random	82.63 ± 15.11	85.22 ± 15.21
✗	1 Random	82.50 ± 16.04	85.10 ± 16.14
✓	5 Random	82.65 ± 14.77	85.25 ± 14.87
✓	Centroid	82.69 ± 15.14	85.28 ± 15.21

图 2 展示了一幅 EUS 图像的分割结果。两种 TextSAM-EUS （自动和手动）变体与真实情况高度吻合。完全监督的 nnUNet 在掩膜边缘略有过度和不足分割，但表现仅比 TextSAM-EUS 略逊一筹。然而，最近的 SAM 扩展却产生了误报区域（SAM-Point、SAMUS-Point、SAMed），在依赖边界框时（SAM-BBox、MedSAM）经常超出目标。AutoSAM 和 AutoSAMUS 减少了人工工作量，但在该图像示例中仍然留有小的间隙或多余的掩膜。

我们的结果证实，添加语言背景能够像在 SAM 中手动绘制几何提示一样有效地指导分割，而后者需要放射学知识。我们还展示了文本提示学习优于其他 SOTA SAM 方法中使用的架构，无论是手动还是自动的。例如，SAMUS 将一个 CNN 附加到 SAM 的 ViT 骨干上

并微调其 39.65M 参数，其性能比只训练 1.69M 参数的 TextSAM-EUS 要差。最终，我们的研究表明，对于胰腺癌 EUS，文本令牌与 ViT 特征网格的对齐比增加外部 CNN 路径更有价值，可能也适用于其他领域。在一项初步消融实验中，将 SAMUS 编码器交换到我们的框架中，在 LoRA 微调下最多达到了 67.53 % DSC，证实了额外的 CNN 并无帮助。总体而言，我们的研究指出，可学习文本提示在直接向 ViT 特征提供高级语义指导方面的优势，而不是依赖于额外 CNN 分支的局部纹理过滤。除了 EUS 中的胰腺肿瘤之外，我们的文本驱动方法可能在更多方面中应用，特别是在注释医学数据集有限的情况下，因为我们使用的 EUS 数据集仅包括 18 名患者。有趣的是，尽管可以通过可学习文本提示和微调省略手动几何提示，但我们的迭代分割精细化模块表明，从基于文本的预测中派生的几何提示仍略微提升了掩膜质量。此外，较低的可训练参数数量表明该模型有潜力在计算资源较低的临床环境中使用。在未来的工作中，我们计划将我们的框架扩展到多类分割，这可能需要进一步调整提示设计。我们还旨在评估 TextSAM 框架在更多应用和成像模态上的表现。

4. 结论

总结起来，我们引入了 TextSAM-EUS ，这是第一个基于文本提示学习的医学图像分割 SAM。通过将提示学习与 BiomedCLIP 和基于 LoRA 的 SAM 微调整合，我们的框架仅通过调整总参数的 0.86 % (1.69 M) 就能学习，且无需任何手动几何提示而提供准确的肿瘤掩膜。在胰腺的公共内镜超声数据库上，TextSAM-EUS 显著优于考虑中的所有不同类别的竞争自动方法。当提供手动提示时，它也能匹配或超越最佳手动提示引导基线。这些结果突显了语言驱动提示对于超声分割的价值，并为可提示基础模型在更广泛的生物医学应用中打开了大门。

致谢

本研究得到了 MEDTEQ+ 联盟、Alpha Tau Medical 和 MITACS 的资助。部分研究是作为 TransMedTech Institute 活动的一部分进行的，部分感谢于 Apogee 加拿大研究卓越基金的财务支持。Y.X. 得到了魁北克卫生研究基金会（FRQS-chercheur boursier Junior 1）和魁北克帕金森基金会 [34] 的支持。P.S. 得到了加拿大自然科学与工程研究委员会（596537）和魁北克自然科学与技术研究基金（B1X-348625）[31] 的支持。T.K. 得到了魁北克自然科学与技术研究基金（B2X-363874）[20] 的支持。C.S.M. 是 Alpha Tau Medical 的顾问。M.K-O. 得到了魁北克卫生研究基金会（FRQS-chercheur boursier Junior 2）的支持。

References

[1] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In European conference on computer vision, pages 1–21. Springer, 2022.

[2] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation, 2023.

[3] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision, pages 205–218. Springer, 2022.

[4] Qinglong Cao, Zhengqin Xu, Yuntian Chen, Chao Ma, and Xiaokang Yang. Domain-controlled prompt learning. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 936–944, 2024.

[5] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes. ArXiv, abs/2305.00035, 2023.

[6] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain?, 2021.

[7] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision, 132(2):581–595, 2024.

[8] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation, 2023.

[9] Shreyank N Gowda and David A Clifton. Cc-sam: Sam with cross-feature attention and context for ultrasound image segmentation. In European Conference on Computer Vision, pages 108–124. Springer, 2024.

[10] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1):1–23, 2021.

[11] Xin Huang, Jin Wang, Yuan Zhang, and Haibo Liu. Cascade segmentation framework for pancreatic tumor in ultrasound images. IEEE Access, 8:126806–126815, 2020.

[12] Ziyi Huang, Hongshan Liu, Haofeng Zhang, Xueshen Li, Haozhe Liu, Fuyong Xing, Andrew Laine, Elsa Angelini, Christine Hendon, and Yu Gan. Push the boundary of sam: A pseudo-label correction framework for medical segmentation, 2023.

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

[14] María Jaramillo, Josué Ruano, Martín Gómez, and Eduardo Romero. Endoscopic ultrasound database of the pancreas. In 16th International Symposium on Medical Information Processing and Analysis, pages 130–135. SPIE, 2020.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International conference on machine learning, pages 4904–4916. PMLR, 2021.

[16] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19113–19122, 2023.

[17] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15190–15200, 2023.

[18] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Muzammal Naseer, Luc Van Gool, and Federico Tombari. Learning to prompt with text only supervision for vision-language models. arXiv preprint arXiv:2401.02418, 2024.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[20] Taha Koleilat. Transformer les soins de santé grâce à des modèles de langage multimodaux universels, efficaces et évolutifs pour l'analyse biomédicale. https://doi.org/10.69777/363874, 2025. Bourses de doctorat en recherche, numéro de demande 363874, Fonds de recherche du Québec (FRQ).

[21] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-sam: Bridging text and image towards universal medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 643–653. Springer, 2024.

[22] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. arXiv preprint arXiv:2409.19483, 2024.

[23] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Biomedcoop: Learning to prompt for biomedical vision-language models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 14766–14776, 2025.

[24] Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 24–34. Springer, 2024.

[25] Yifan Liu, Lanfen Jin, Minhao Xie, Kai Wang, and Xiahai Xu. Semi-supervised segmentation of pancreatic tumor in ultrasound images with multi-task consistency learning. Medical Image Analysis, 70:101998, 2021.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[27] Yu Lu, Jin Xue, Minhao Xie, Kai Wang, and Xiahai Xu. A deep learning framework for pancreatic tumor segmentation in ultrasound images. Ultrasound in Medicine & Biology, 47(4):968–982, 2021.

[28] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. Nature Communications, 15(1), 2024.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[30] Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. arXiv preprint arXiv:2306.06370, 2023.

[31] Pascal Spiegler. Planification automatique de la trajectoire des Électrodes pour la chirurgie de stimulation cérébrale profonde connectomique. https://doi.org/10.69777/348625, 2025. Bourse de maîtrise en recherche, numéro de demande 348625, Fonds de recherche du Québec (FRQ).

[32] Pascal Spiegler, Amirhossein Rasoulian, and Yiming Xiao. Weakly supervised intracranial hemorrhage segmentation with yolo and an uncertainty rectified segment anything model. In MICCAI Challenge on Ischemic Stroke Lesion Segmentation, pages 12–21. Springer, 2024.

[33] Qi Wu, Yuyao Zhang, and Marawan Elbatel. Self-prompting large vision models for few-shot medical image segmentation. In MICCAI workshop on domain adaptation and representation transfer, pages 156–167. Springer, 2023.

[34] Yiming Xiao. Améliorer le prognostic neurochirurgical avec l'imagerie multimodale et à la connectivité cérébrale. https://doi.org/10.69777/330745, 2024. Chercheurs-boursiers, numéro de demande 330745, Fonds de recherche du Québec (FRQ).

[35] Yan Xu, Rixiang Quan, Weiting Xu, Yi Huang, Xiaolong Chen, and Fengyuan Liu. Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches. Bioengineering, 11(10):1034, 2024.

[36] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785, 2023.

[37] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930, 2021.

[38] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024.

[39] Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, et al. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. arXiv preprint arXiv:2405.12971, 2024.

[40] Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A comprehensive survey. arXiv preprint arXiv:2312.07353, 2023.

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In CVPR, pages 16816–16825, 2022.

[42] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. IJCV, 130(9):2337–2348, 2022.