arXiv:2602.09252v1 [cs.CV] 9 Feb 2026

基础模型助力的视觉语言模型引导的手术图像分割迭代细化

Ange Lou [*1] **Yamin Li** [*1] **Qi Chang** [2] **Nan Xi** [3] **Luyuan Xie** [4] **Zichao Li** [5] **Tianyu Luan** [6]

摘要

手术图像分割对于机器人辅助手术和术中引导至关重要。然而，现有方法受限于预定义类别，只能生成一次性预测，无法进行自适应细化，也缺乏临床医生交互机制。我们提出 IR-SIS，这是一种接受自然语言描述的手术图像分割迭代细化系统。IR-SIS 利用经过微调的 SAM3 进行初始分割，借助视觉语言模型检测手术器械并评估分割质量，并采用智能体工作流自适应地选择细化策略。该系统支持通过自然语言反馈实现临床医生参与式交互。我们还基于 EndoVis2017 和 EndoVis2018 基准数据集构建了一个带有多粒度语言标注的数据集。实验表明，该方法在域内和分布外数据上均达到最先进的性能，临床医生交互还能进一步提升效果。我们的工作首次建立了具备自适应自我细化能力的语言驱动手术分割框架。

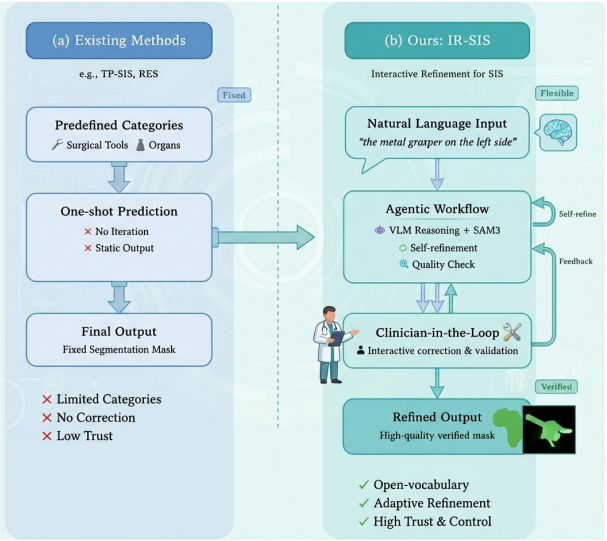

图 *1.* 现有基于语言的手术分割方法面临两个关键局限：a）它们受限于预定义器械类别（例如 TP-SIS），只能进行一次性预测，无法自适应细化或与临床医生交互。（b）我们的方法 IR-SIS 支持灵活的自然语言描述，采用智能体工作流进行自适应自我细化，并支持临床医生参与式交互以提升质量。

1. 引言

手术图像分割是计算机辅助介入系统的一项基础能力。对手术器械和组织进行准确分割，能够支持手术导航、机器人自动化和术后分析。然而，真实的手术环境带来了诸多挑战：新的器械不断被引入，不同的手术操作需要分割不同的目标，而临床工作流程既要求精确性，也要求适应性。这些需求要求分割系统能够通过灵活的用户交互，应对多样且不断变化的器械类型。

现有的手术图像分割方法在两个根本方面无法满足这些要求。传统方法基于预定义器械类别进行训练，因此在遇到训练数据中未包含的器械时，性能会下降。鉴于手术工具快速发展，这种情况十分常见（Zhou 等，2023）。近期的 TP-SIS 等基于语言的方法（Zhou 等，2023）通过接受文本提示而非类别ID，展现了进步。然而，TP-SIS 仍受限于固定的器械类型词汇表（例如“镊子”“剪刀”），无法理解“左侧的器械”或“抓取

[1]美国田纳西州纳什维尔范德堡大学 [2]美国宾夕法尼亚州大学公园宾夕法尼亚州立大学 [3]美国弗吉尼亚州里士满弗吉尼亚联邦大学 [4]中国北京北京大学 [5]美国加利福尼亚州圣迭戈加利福尼亚大学圣迭戈分校 [6]中国广州华南理工大学自动化科学与工程学院。通讯作者：Tianyu Luan <luanty.henry@gmail.com>。

预印本。*2026*年*2*月*11*日。

组织附近的工具”等自由形式描述。更重要的是，传统方法和基于语言的方法都采用单次预测：模型在一次前向传播中生成分割掩码，却没有质量评估、迭代细化或纳入临床医生反馈的机制。初始分割出现错误时，无论是由器械边界模糊、遮挡还是异常光照造成，这些方法都无法进行纠正。缺乏自适应细化和人在回路中的交互，限制了它们在临床环境中的可靠性，而分割错误可能会造成后续影响。

我们观察到，人类专家不会一次性完成手术图像标注。相反，他们会遵循一个迭代流程：初始标注、质量评估、针对性纠正和确认。这一观察引出了我们的核心观点：分割系统应模拟这种迭代工作流，而非依赖单次预测。视觉语言模型（VLM）如今具备强大的视觉理解与推理能力，远超简单分类的范畴。VLM 可以检测图像中的器械，用边界框对其进行定位，并评估给定的分割掩码是否充分覆盖检测到的目标。将 VLM 视为质量评估器后，我们便能构建一个智能体工作流，根据量化的质量指标迭代细化分割结果。这种架构还为临床医生参与其中提供了自然的交互接口：当 VLM 的自动评估不够充分时，临床医生可以提供有针对性的反馈，指导后续的细化迭代。

基于这一观点，我们提出了 IR-SIS（手术图像分割的迭代细化），这是一种视觉语言模型引导框架，可将手术图像分割从单次预测转变为自适应迭代细化。IR-SIS 包含四个组件：(1) 一个基于 SAM 的分割器（Kirillov 等人，2023；Ravi 等人，2024），该分割器在我们的多层级语言标注数据集上进行了微调，用于执行基于语言的初始分割；(2) 一个视觉语言模型器械检测器，用于识别所有手术器械并输出边界框；(3) 一个质量评估器，用于计算掩码覆盖率和掩码-边界框重叠度指标；以及 (4) 一个智能体细化工作流，根据质量分数在信任初始结果和多器械策略之间进行选择。当质量指标低于阈值时，系统会使用检测到的边界框作为提示，对各个器械重新分割，并通过迭代改进结果。临床医生可以在任意迭代中介入并提供纠正性反馈。

我们的贡献如下：

- **基于语言的手术分割系统**。我们提出了一种分割系统，可接受对目标器械的自然语言描述，使临床医生能够灵活指定分割目标，而不必从预定义类别中选择。该系统还可适应多样的临床词汇和新型器械，无需重新训练。

- **智能体迭代细化工作流**。 我们设计了一种由视觉语言模型引导的工作流，用于评估分割质量并进行自适应细化，将范式从单次预测升级为智能迭代。该工作流支持临床医生参与式交互，使人工反馈能够在自动评估不足时指导细化。

- **多层级语言标注数据集**。 我们构建了一个数据集，其中包含三个粒度级别（通用、类别和特定）的标注，覆盖 EndoVis2017 和 EndoVis2018 基准数据集。该数据集支持训练能够理解不同特异性级别查询的模型，从“手术器械”到“双极电凝钳”。

- **全面的实验验证**。 我们在域内和分布外数据上评估了 IR-SIS，结果表明，智能体工作流能够提升分割质量，临床医生交互可带来额外收益，而基于我们数据集训练的模型对未见过的器械展现出更强的泛化能力。

2. 相关工作

2.1. 手术器械分割

深度学习已成为手术器械分割的主流方法。早期基于U-Net及其变体的方法在En- doVis基准数据集上取得了良好性能（Shvets等，2018年；Allan等，2019年）。ISINet（Gonz´alez等，2020年）引入了结合时序一致性建模的实例级分割，而TraSeTR（Zhao等，2022年）在Transformer框架中利用跟踪线索，以改进实例级预测。Segment Anything模型（SAM）（Kir- illov等，2023）展现了强大的零样本分割能力，推动了其在医学影像领域的适配。MedSAM（Ma等，2024）在涵盖多种成像模态的大规模医学数据上对SAM进行了微调。SurgicalSAM（Yue等，2024）和AdaptiveSAM（Paranjape等，2024）专门针对手术场景分割，并采用高效的微调策略。然而，这些方法需要明确的几何提示（点、框），且缺乏理解目标器械自然语言描述的能力。

2.2. 语言引导的图像分割

对比语言-图像预训练（CLIP）（Radford 等，2021）使视觉语言模型能够理解图像与文本之间的语义关系。CRIS（Wang 等，2022）和 CLIPSeg（L¨uddecke 与 Ecker,

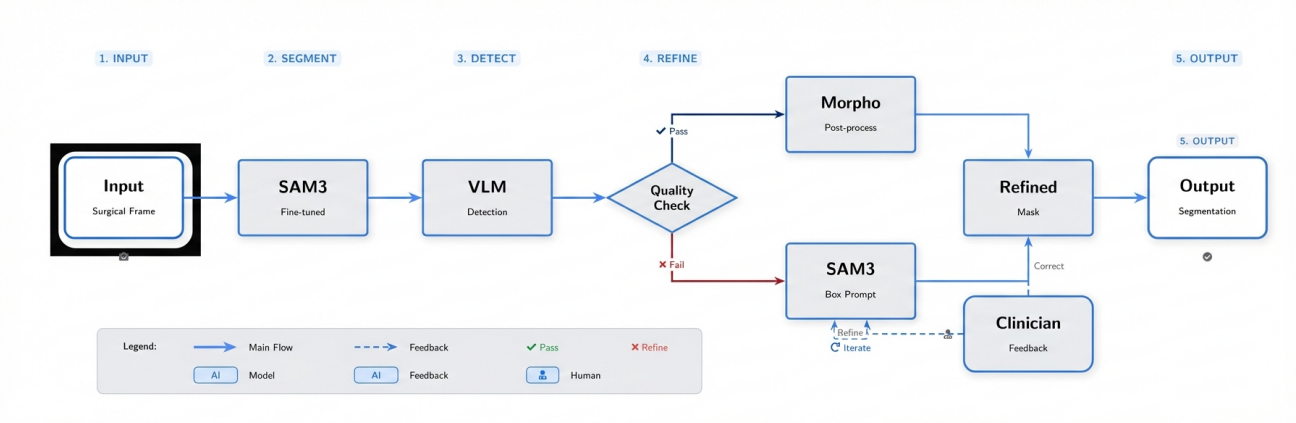

图 2. IR-SIS 流程概览。我们首先在自有数据集上微调 SAM3，使其对输入的手术图像进行初始分割。随后，使用视觉语言模型检测所有手术器械。接着，根据掩码覆盖率和掩膜框重叠率评估分割质量：如果质量符合标准，则仅进行形态学后处理；否则，使用框提示对每个器械重新分割，并进行迭代细化。此外，临床医生可以在迭代过程中提供反馈，以指导细化方向，最终获得细化后的分割结果。

2022）将 CLIP 的语义知识迁移至像素级分割，而 LAVT（Yang 等，2022）则将语言信息直接融入视觉 Transformer 特征。在手术领域，TP-SIS（Zhou 等，2023）率先采用混合提示机制实现文本提示式器械分割。然而，TP-SIS 及相关方法仍受限于固定类别提示，且只能生成单次预测，无法进行迭代细化。交互式分割方法（Marinov 等，2024；Zhu 与 Wu，2024）通过纳入用户反馈实现迭代改进，但每次迭代都需要人工干预。我们的 IR-SIS 与以往工作有三点不同：（1）接受灵活的自然语言描述，而非固定类别提示；（2）采用视觉语言模型引导的智能体工作流，实现自主质量评估与自适应细化；（3）支持可选的临床医生参与式交互，以便在需要时进行引导式改进。

3. 方法

3.1. 问题形式化

我们研究基于语言的手术图像分割任务。给定一幅手术图像 $I \in \mathbb{R}^{H\times W\times 3}$ 和一个描述目标器械的自然语言查询 q，目标是预测一个二值分割掩膜 $M \in \{0,1\}^{H\times W}$，以识别属于查询对象的所有像素。与只能处理固定器械类别的传统手术分割方法不同，我们的方法支持多种粒度层级的自由形式文本查询，既可以是一般性描述（“手术器械”），也可以是具体标识（“双极电凝钳”）。我们将目标形式化为：

$$\hat{M} = \arg\max_{M} p(M|I, q; \theta) \tag{1}$$

其中 θ 表示模型参数。为实现这一目标，同时支持自适应细化和临床医生交互，我们提出了 IR-SIS，这是一种由视觉语言模型引导的迭代细化系统。

3.2. 方法概述

如图 2所示，IR-SIS 包含四个阶段，将手术图像分割从一次性预测转变为自适应迭代细化。首先，我们基于微调后的 SAM 架构构建语言引导分割器（Kirillov 等人，2023；Ravi 等人，2024），根据图像和文本查询生成初始分割掩码 M_0。其次，视觉语言模型（VLM）检测器（Bai 等人，2023）识别图像中所有可见的手术器械，并输出其边界框 $B = \{b_1, \ldots, b_n\}$。第三，质量评估器计算掩码覆盖率和掩码-边界框重叠度指标，以评估初始分割是否充分覆盖目标器械。第四，智能体细化工作流利用这些质量评分选择相应策略：如果质量足够，系统便进行形态学后处理并输出结果；否则，系统使用框提示对低质量区域进行迭代重分割，直至达到质量阈值或达到最大迭代次数。临床医生可以在任意迭代阶段提供反馈，以指导细化方向。下面我们将详细介绍各个组件。

3.3. IR-SIS 流程

语言引导分割器。 现有的外科图像分割方法，例如 TP-SIS（Zhou 等，2023），仅支持固定词汇提示，无法泛化到自由形式语言描述或新型器械类型。为解决这一局限，我们对基于 SAM 的架

构（Kirillov 等，2023；Ravi 等，2024）进行微调，采用分层学习率策略，在适应外科图像分割的同时保留预训练特征。

视觉骨干网络采用较低的学习率并配合层衰减，以保留预训练的视觉表征。文本编码器保持冻结，以利用其预训练的语言理解能力。Transformer 解码器采用较高的学习率来学习分割任务。此配置使模型能够适应手术领域，同时避免灾难性遗忘通用视觉知识。

为增强模型在不同查询粒度下的泛化能力，我们使用多级提示进行训练。每个已标注器械都会生成三个特异性级别不同的训练样本：级别 0 为通用查询（例如“手术器械”），级别 1 为类别级查询（例如“镊子”），级别 2 为特定器械查询（例如“ 双极镊子”）。分割器会生成初始掩码：

$$M_0 = f_{\text{SAM}}(I, q; \theta) \tag{2}$$

其中， θ 表示微调参数。

视觉语言模型器械检测器。初始分割可能会漏掉器械，或包含查询目标之外的多个对象。为了全面了解图像中的所有器械，我们使用视觉语言模型（Bai 等人，2023）检测图像中可见的手术器械。给定图像 I和检测提示 p_{detect}，视觉语言模型会输出所有检测到的器械的边界框：

$$B = \{b_i\}_{i=1}^{n} = \text{VLM}(I, p_{\text{detect}}) \tag{3}$$

其中， n 为检测到的器械数量。这些边界框有两个用途：既提供参考区域以进行质量评估，也在需要时为迭代细化提供提示。

质量评估器。 根据检测结果，我们量化初始分割的质量，以确定是否需要细化。我们定义了两个互补指标。掩码覆盖率 用于衡量检测到的器械区域中被分割结果覆盖的比例：

$$C = \frac{|M \cap \bigcup_i b_i|}{|\bigcup_i b_i|} \tag{4}$$

掩码-边界框重叠度 用于衡量分割结果与目标边界框的对齐程度：

$$O = \frac{|M \cap b_{\text{target}}|}{|b_{\text{target}}|} \tag{5}$$

质量指标综合了以下指标：

$$\mathcal{S} = H(C > \tau_c) \wedge H(O > \tau_o) \tag{6}$$

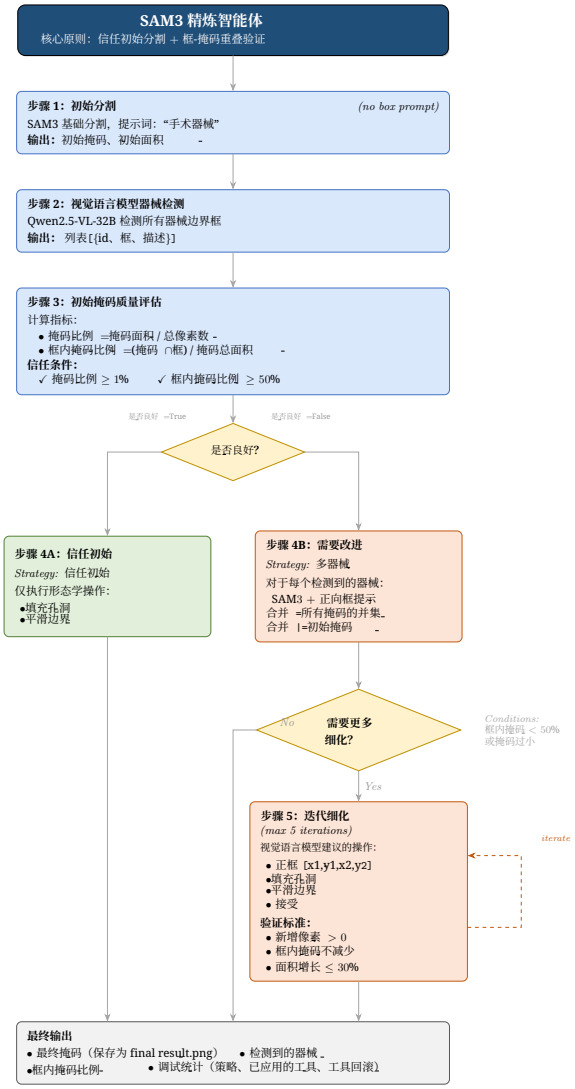

图 3. SAM3 精炼智能体的详细工作流程。该智能体根据掩码-边界框重叠质量，自适应地在信任初始结果策略和多器械策略之间进行选择。

其中， τ_c 和 τ_o 是根据验证集调优的阈值。$H(\cdot)$是一个阶跃函数，当数值高于阈值时输出 1，否则输出 0，起到简单的质量开关作用。

智能体细化工作流。 在涉及遮挡、多种器械或低对比度的复杂手术场景中，单次预测往往表现不佳。根据质量评估，我们的智能体工作流会自适应地选择适当的细化策略。图 3 展示了该工作流的伪代码级流程。

当质量指标表明初始分割已足够好时（$\mathcal{S} = 1$），系统仅进行形态学后处理。开运算会去除小型

表 *1.* 我们的方法与其他最先进方法在 EndoVis2018 数据集上的比较 .

方法	Ch IoU	BF	PF	LND	SI	CA	MCS	UP	mc IoU
TernausNet-11（Iglovikov 与 Shvets，2018年）	46.22	44.20	4.67	0.00	0.00	0.00	50.44	0.00	14.19
MF-TAPNet（Jin 等人，2019年）	67.87	69.23	6.10	11.68	14.00	0.91	70.24	0.57	24.68
ISINet（Gonz´alez 等人，2020年）	73.03	73.83	48.61	30.98	37.68	0.00	88.16	2.16	40.21
TraSeTR（Zhao 等人，2022年）	76.20	76.30	53.30	46.50	40.60	13.90	86.30	17.50	47.77
S3Net（Xue 等人，2022年）	75.81	77.22	50.87	19.83	50.59	0.00	92.12	7.44	42.58
MATIS（Ayobi 等人，2023）	84.26	83.52	41.90	66.18	70.57	0.00	**92.96**	23.13	54.04
CRIS（Wang 等人，2022年）	74.10	73.58	58.20	47.64	72.14	4.56	45.99	20.18	46.04
CLIPSeg（L¨uddecke & Ecker，2022年）	74.95	67.25	39.59	36.72	47.27	2.92	79.96	4.22	39.7
TP-SIS(448)（Zhou 等，2023）	82.67	81.53	70.18	71.54	90.58	21.46	65.57	57.51	65.48
TP-SIS(896)（Zhou 等，2023）	84.92	84.28	73.18	78.88	92.20	23.73	66.67	39.12	65.44
Ours	**85.44**	**87.99**	**82.69**	**49.65**	**86.41**	**31.12**	92.86	**79.39**	**72.87**

表 *2.* Kvasir-Instrument 数据集上的视觉语言模型比较

视觉语言模型	平均交并比	改进
基线（微调后的 SAM3）	71.08	N/A
Qwen2.5-VL-32B-Instruct	77.95	+6.87
Qwen3-VL-32B-Instruct（我们的方法）	78.90	+7.82

噪声伪影，闭运算则会填充孔洞： $M_{\text{final}} = \text{Morph}(M_0)$。当质量不足时（$\mathcal{S} = 0$），系统会识别出哪些边界框中的分割质量较差，并使用框提示对这些区域重新分割。该过程会持续迭代，直到达到质量阈值或达到最大迭代次数：

$$M^{(t+1)} = \begin{cases} \text{Morph}(M^{(t)}) & \text{if } \mathcal{S}^{(t)} = 1 \\ \text{SAM}(I, b_{\text{low}}) \cup M^{(t)}_{\text{good}} & \text{if } \mathcal{S}^{(t)} = 0 \end{cases} \tag{7}$$

其中 b_{low} 表示分割质量较低的边界框， $M^{(t)}_{\text{良好}}$ 表示从迭代 t中保留的高质量部分。

该工作流支持临床医生在任意迭代中进行干预，提供三种交互模式：(1) 使用框提示指定感兴趣区域或遗漏的器械；(2) 使用语言描述补充或更正查询；(3) 使用参考标注引导分割。临床医生反馈优先于自动视觉语言模型评估，从而确保临床可控性。

3.4. 训练

损失函数。 我们使用复合损失函数训练语言引导分割器：

$$\mathcal{L}_{\text{total}} = \lambda_{\text{mask}} \cdot \mathcal{L}_{\text{mask}} + \lambda_{\text{dice}} \cdot \mathcal{L}_{\text{dice}} + \lambda_{\text{ce}} \cdot \mathcal{L}_{\text{ce}} + \lambda_{\text{presence}} \cdot \mathcal{L}_{\text{presence}} \tag{8}$$

掩码损失 $\mathcal{L}_{\text{mask}}$ 是一种焦点损失（α=0.25， γ=2），用于处理前景/背景不平衡。Dice 损失 $\mathcal{L}_{\text{dice}}$可改善小目标分割。交叉熵损失 $\mathcal{L}_{\text{ce}}$ 用于监督查询-目标匹配分类。存在性损失 $\mathcal{L}_{\text{presence}}$ 用于预测图像中是否存在被查询的目标。在我们的实验中，我们将 λ_{mask}、 λ_{dice}、λ_{ce} 和 $\lambda_{\text{presence}}$ 分别设为 5、1、2、2。

分层学习率策略。 不同模型组件采用不同的学习率，以平衡适配与预训练知识的保留。Transformer 解码器使用学习率 8×10^{-5} 进行任务特定学习。视觉骨干网络使用 2.5×10^{-5}，并采用层衰减（0.98）来保留预训练特征。文本编码器保持冻结，以保留语言理解能力。

匹配策略。 我们采用匈牙利匹配结合一对多匹配（top-k=4，权重=2.0），以平衡正负样本并提高训练稳定性。

数据增强。 我们采用领域特定的数据增强，以提高模型对手术成像条件的鲁棒性：使用颜色抖动应对颜色变化，使用伽马校正（$\gamma \in [0.8, 1.2]$）实现曝光自适应，并通过高光增强和随机阴影增强模拟手术场景光照条件。

质量评估阈值。 我们在验证集上调整覆盖率阈值τ_c 和重叠阈值 τ_o，以平衡质量评估中的精确率与召回率。形态学操作使用的核尺寸根据训练数据中典型的器械尺寸进行选择。

4. 实验

我们在域内和分布外手术图像分割任务上评估 IR-SIS。实验结果证明了所提出的智能体工作流的有效性，以及临床医生参与式交互的益处。

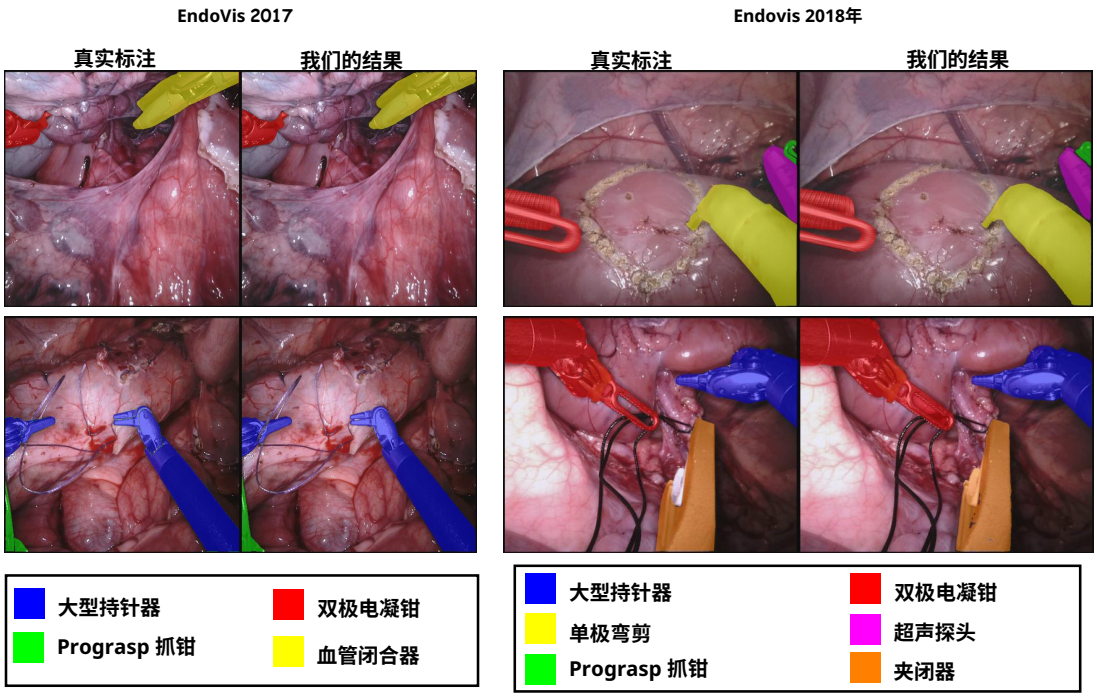

图 *4.* EndoVis2017 和 EndoVis2018 数据集上的文本提示分割结果

4.1. 数据集

训练数据。我们使用来自 EndoVis2017 和 EndoVis2018 挑战赛的合并数据集训练模型。EndoVis2017 包含来自 8 段机器人手术序列的 1,778 张图像，涵盖多种器械类型。EndoVis2018 提供了来自 15 段手术视频的 1,630 张图像，器械外观多样。两个数据集均采用我们的多层级语言标注（L0-L2）进行标注，经多级提示扩增 3 倍后，共得到 3,067 张训练图像和 19,494 条标注。

评估数据。对于域内评估，我们使用 EndoVis2017 和 EndoVis2018 的测试集，共包含 341 张图像。对于分布外评估，我们使用 Kvasir-Instrument，该数据集包含胃肠内窥镜检查图像，其器械类型和视觉特征在训练期间未曾出现。

4.2. 评估指标

我们采用两种标准分割指标。**Dice 系数（DSC）**用于衡量预测掩码与真实掩码之间的重叠程度，计算方式为 $\text{DSC} = 2|A \cap B|/(|A| + |B|)$。**交并比（IoU）**则从另一个角度衡量分割质量，定义为 $\text{IoU} = |A \cap B|/|A \cup B|$。这两项指标的取值范围均为 0 到 1，数值越高表示分割效果越好。

4.3. 实施细节

模型配置。我们以 SAM3 作为基础分割模型进行微调，输入分辨率为 1008×1008。对于视觉语言模型组件，我们使用 Qwen-2.5-VL-32B- instructas thedefault instrument detector，与 GPT-4o 及其他模型进行比较。

训练设置。我们使用 4 个 NVIDIA A6000 GPU（每个 48GB）进行训练，总批量大小为 64（每个 GPU 为 16），并采用 FP16 混合精度。我们采用分层学习率：8×10^{-5}用于 Transformer 解码器，2.5×10^{-5}用于视觉骨干网络，并设置 0.98 的层衰减，同时冻结文本编码器。总损失由掩码焦点损失（权重=50, α=0.25, γ=2）、Dice 损失（权重=10）、分类焦点损失（权重=20）和存在性损失（权重=20）组成。训练采用 90 个 epoch 的训练计划：30 个 epoch 预热、30 个 epoch 衰减，以及 30 个 epoch 冷却。

数据增强。我们采用颜色抖动（p=0.8）、伽马校正（p=0.5）、镜面高光模拟（p=0.3）和随机阴影（p=0.3），以增强模型对手术成像变化的鲁棒性。

基线方法。我们将以下方法作为对比：(1) 未进行微调的 SAM3；(2) 在我们的数据集上进行微调、但不使用智能体工作流的 SAM3；(3) 在零样本设置下进行评估的 TP-SIS。

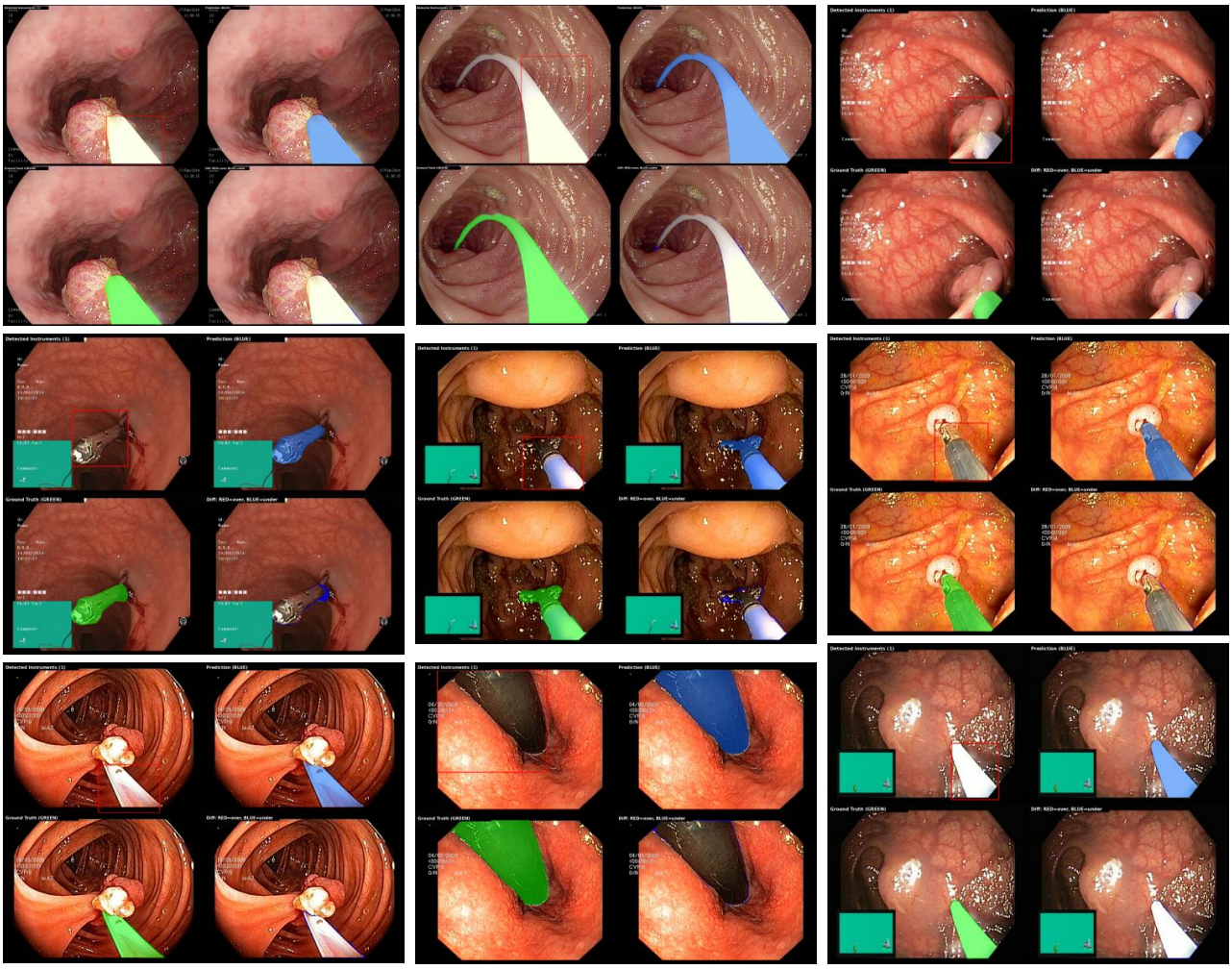

图 *5.* Kvasir-Instrument 数据集上的文本提示分割结果

4.4. 实验结果

主要比较。 我们在域内和分布外数据上，将 IR-SIS 与基线方法进行比较。在域内 EndoVis 测试集上（表 1），我们的方法相比 SAM3 原始模型和 TP-SIS 取得了显著改进。在我们的多层级标注数据集上进行微调为模型提供了坚实基础，而智能体细化工作流则通过自适应地纠正初始预测，进一步提升了分割质量。

在分布外的 Kvasir-Instrument 数据集上（表 2 和图 5），与基线方法相比，IR-SIS 展现出更强的泛化能力。在这种情境下，视觉语言模型引导的迭代细化尤为有效，因为它无需特定数据集训练即可检测并纠正分割错误。这验证了我们的设计选择：将微调后的分割模型与自适应细化机制相结合。

视觉语言模型比较。 我们比较了不同视觉语言模型作为表格中的智能体工作流器械检测器 2。GPT-4o（API）取得了最高的检测准确率，但会产生 API 费用和延迟。Qwen-2.5-VL-32B-instruct 在本地部署时表现不俗，在准确率与易用性之间实现了实用的平衡。视觉语言模型的选择会影响器械检测准确率，进而影响最终的分割质量；器械检测准确率则为迭代细化过程提供指引。

语言多样性演示。 IR-SIS 支持不同粒度层级的灵活自然语言查询。如图 6所示，对于同一张手术图像，用户可以分割“手术器械”（所有器械）、“镊子”（某一特定类别）、“双极电凝钳”（某一种特定器械），甚至可以使用“左侧的器械”这类空间描述。这种灵活性让临床医生能够自然地进行交互，因为他们可能会根据自身专业知识和临床情境，以不同方式描述器械。

消融研究。 我们进行消融实验以

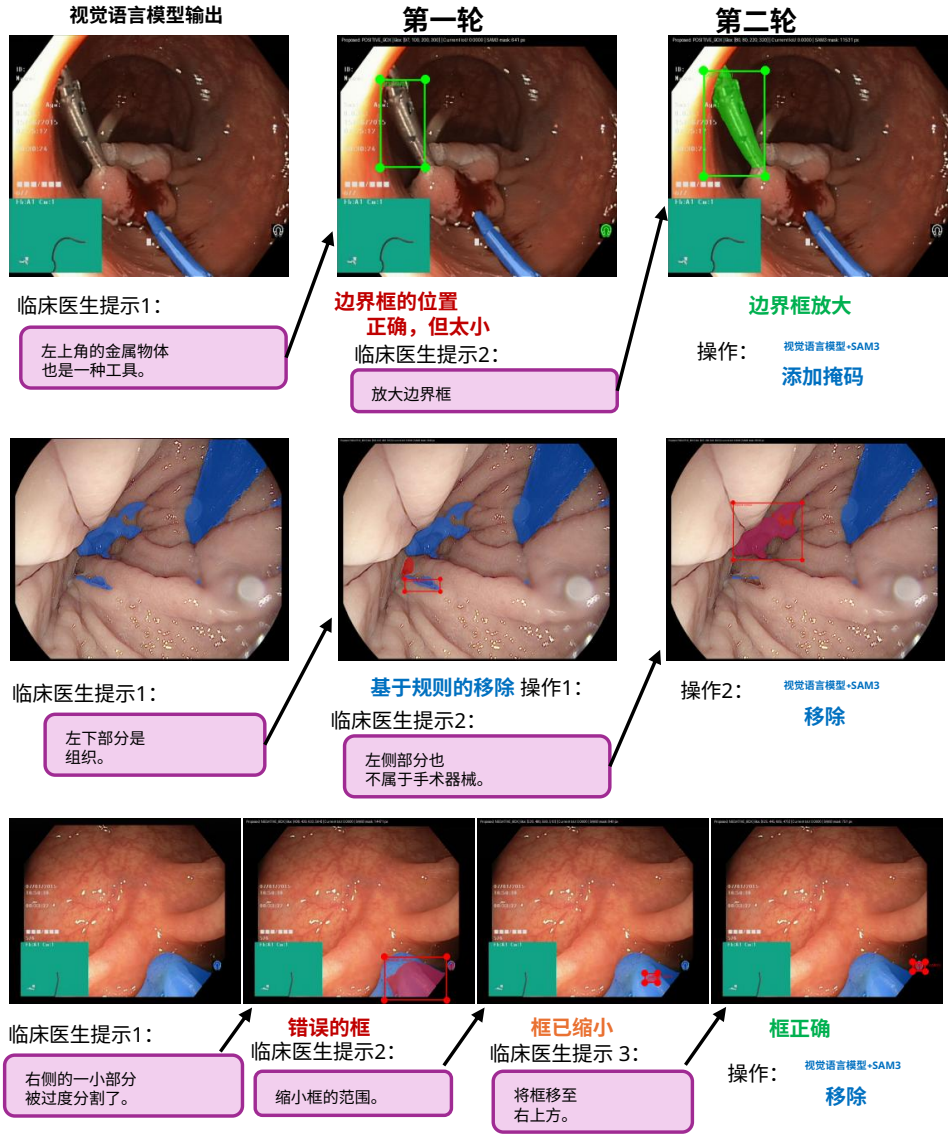

图6. 临床医生参与闭环示例。

验证关键设计选择。我们在图 6 中评估临床医生反馈对分割质量的影响。在细化过程中，临床医生提供纠正提示后，分割准确率相比仅由视觉语言模型进行的细化有所提升。这表明，我们的系统能够有效利用人类专业知识，处理自动细化无法胜任的棘手案例。

5. 结论

我们提出IR-SIS，一种视觉语言模型引导框架，将手术图像分割从一次性预测转变为自适应迭代细化。通过将在多层级标注上微调的基于语言的分割器与利用视觉语言模型检测的器械的智能体工作流相结合以及质量指标，IR-SIS克服了现有方法中预定义类别的局限以及缺乏自我纠正的问题。实验表明，通过迭代细化可提升分割质量，临床医生参与式交互可带来额外增益，并且对未见过的器械具有较强的泛化能力。

参考文献

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Du- ber, R., Joos, S., Maier-Hein, L. 等。2017 年机器人器械分割挑战赛。*arXiv* 预印本 *arXiv:1902.06426*，2019年。

Ayobi, N., P´erez-Rond´on, A., Rodr´ıguez, S. 和 Arbel´aez, P. MATIS：用于手术器械分割的掩码注意力变换器

。*2023* 年 *IEEE* 第*20*届国际生物医学成像研讨会（*ISBI*），第 1–5 页。IEEE，2023 年。

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl：一种功能多样的视觉语言模型，可用于理解、定位、文本识读等任务。*arXiv* 预印本 *arXiv:2308.12966*，2023。

Gonz´alez, C., Bravo-S´anchez, L., and Arbelaez, P. ISINet：一种用于手术器械分割的基于实例的方法。载于国际医学图像计算与计算机辅助干预会议（*MIC-CAI*），第 595–605 页。施普林格，2020年。

Iglovikov, V. and Shvets, A. TernausNet：一种采用在 ImageNet 上预训练的 VGG11 编码器的 U-Net，用于图像分割。*arXiv* 预印本 *arXiv:1801.05746*，2018年。

Jin, Y., Cheng, K., Dou, Q., 和 Heng, P.-A. 利用运动流中的时间先验进行微创手术视频中的器械分割。在医学图像计算与计算机辅助手术国际会议，第 440–448 页。施普林格，2019年。

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Doll´ar, P., 和 Girshick, R. Segment Anything。在 *IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，第 4015–4026 页，2023。

L¨uddecke, T. 和 Ecker, A. 使用文本和图像提示进行图像分割。在*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第 7086–7096 页，2022。

Ma, J.、He, Y.、Li, F.、Han, L.、You, C.与Wang, B. 医学图像中的万物分割。《自然-通讯》，15(1):654，2024。

Marinov, Z.、J¨ager, P. F.、Egger, J.、Kleesiek, J. 与Stiefelhagen, R. 医学图像的深度交互式分割：系统综述与分类体系。《*IEEE*模式分析与机器智能汇刊》，46(12):10998–11018，2024。

Paranjape, J. N.、Nair, N. G.、Sikder, S.、Vedula, S. S. 与Patel, V. M. AdaptiveSAM：面向手术场景分割的SAM 高效微调。*arXiv*预印本 *arXiv:2308.03726*，2024。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等. 基于自然语言监督学习可迁移视觉模型. 载于国际机器学习会议（*ICML*），第 8748–8763 页. PMLR，2021.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., R¨adle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Doll´ar, P., 和 Feichtenhofer, C. SAM 2：对图像和视频中的任何对象进行分割. *arXiv* 预印本 *arXiv:2408.00714*，2024.

Shvets, A.A.、Rakhlin, A.、Kalinin, A. A.与 Iglovikov, V. I. 基于深度学习的机器人辅助手术自动器械分割。*arXiv*预印本 *arXiv:1803.01207*，2018年。

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M. 和 Liu, T. Cris：CLIP驱动的指代表达图像分割。载于 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第 11686–11695 页，2022年。

Xue, Z., Zhou, Y. 和 Du, P. S3net：用于少样本高光谱图像分类的光谱—空间孪生网络。*IEEE*地球科学与遥感汇刊，60：1–19，2022年。

Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H. 和 Torr, P. H. Lavt：用于指代表达图像分割的语言感知视觉 Transformer。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第 18155–18165 页，2022年。

Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., and Wang, Z. SurgicalSAM：高效的类别提示式手术器械分割。载于*AAAI* 人工智能会议论文集，第 38 卷，第 6890–6898 页，2024。

Zhao, Z., Jin, Y., Gao, X., Dou, Q., and Heng, P.-A. TraSeTR：一种用于机器人手术中实例级器械分割的、带有对比查询的跟踪到分割Transformer。载于*IEEE* 国际机器人与自动化会议（*ICRA*），第 11186–11193 页。IEEE，2022年。

Zhou, Z., Alabi, O., Wei, M., Vercauteren, T., and Shi, M. 基于视觉语言模型的文本提示式手术器械分割。载于神经信息处理系统进展（*NeurIPS*），第 36 卷，2023。

Zhu, J. 和 Wu, J. Meduhip：迈向人在回路中的医学分割。发表于 医学影像计算与计算机辅助介入国际会议（*MICCAI*），2024。arXiv 预印本 arXiv:2408.01620。