arXiv:2603.01108v1 [cs.CV] 1 Mar 2026

GroundedSurg：用于语言条件手术器械分割的多手术流程基准

Tajamul Ashraf[1 4 †]、Abrar Ul Riyaz*[4]、Wasif Tak*[2 ‡]、Tavaheed Tariq*[4]、Sonia Yadav[4]、、Moloud Abdar[3]、和 JanibulBashir[4]

[1] 沙特阿拉伯阿卜杜拉国王科技大学（KAUST）[2] 印度塔帕尔工程技术学院[3] 澳大利亚昆士兰大学[4] 印度斯利那加国家理工学院 Gaash研究实验室

摘要。 临床可靠的手术场景感知对于推进智能、具备情境感知能力的术中辅助至关重要，例如器械交接引导、碰撞规避和工作流程感知型机器人辅助。现有手术器械基准主要评估类别级分割，要求模型检测所有预定义器械类别的实例。然而，现实临床决策往往需要根据器械的功能角色、空间关系或解剖结构交互能力，确定指代的特定器械实例，而这些能力并未纳入当前的评估范式。我们推出 **GroundedSurg**，这是首个语言条件的实例级手术定位基准。每个样本都将一张手术图像与一段指向单个器械的自然语言描述配对，并附有结构化空间定位标注，包括边界框和点级锚点。该数据集涵盖眼科、腹腔镜、机器人及开放手术流程，包含多种器械类型、成像条件和手术复杂程度。通过联合评估语言指代解析和像素级定位，GroundedSurg 可系统且真实地评估视觉语言模型在临床真实多器械场景中的表现。大量实验表明，现代分割模型和视觉语言模型之间存在显著的性能差距，凸显了手术人工智能系统亟需具备以临床为基础的视觉语言推理能力。代码和数据已公开发布于
https://github.com/gaash-lab/GroundedSurg

关键词： 手术器械分割 · 视觉语言定位 · 手术场景理解 · 医学图像基准

† 通讯作者：tajamul.ashraf@kaust.edu.sa* 贡献相同

‡ 在斯利那加国家理工学院 Gaash 实验室实习期间完成的工作

表 1：不同手术工具数据集在定位能力方面的比较。我们的基准支持多手术流程、语言条件、显式空间定位、实例级消歧和基于提示的评估。

数据集	领域	多手术流程	语			
CaDIS [8]	白内障	✗	✗	✗	✗	✗
AutoLaparo [30]	腹腔镜	✗	✗	✗	✗	✗
SISVE [32]	胃切除术	✗	✗	✓	✓	✗
EndoVis [26]	机器人	✗	✗	✗	✗	✗
CholecSeg8k [10]	胆囊切除术	✗	✗	✗	✗	✗
CholecInstanceSeg [1]	胆囊切除术	✗	✗	✗	✓	✗
鲁棒医学图像分割 [22]	MIS	✗	✗	✗	✗	✗
InSeg1 [6]	眼科的	✗	✗	✓	✓	✗
InSeg2 [6]	眼科的	✗	✗	✓	✓	✗
GroundedSurg（我们的）		✓	✓	✓	✓	✓

1 引言

对手术场景的准确解读，传统上被表述为对预定义器械类别进行分割和识别 [1,2,15]。这类类别层面的表述推动了多项重要下游应用，包括工作流程分析、技能评估和情境感知辅助系统 [23, 28, 29]。然而，手术场景理解并不止于类别识别。在真实的手术环境中，多个外观相似的器械往往同时出现在同一视野中。它们的临床相关性不仅取决于类别，还取决于功能角色、空间配置以及与周围解剖结构的相互作用 [6, 26]。因此，要区分正在主动分离组织的器械与另一把同类型但处于闲置或牵拉状态的器械，就必须根据上下文确定所指的具体实例。

当前评估范式并未评估这一能力 [6,8,23]。类别级器械分割基准衡量识别性能，但当多个候选对象同时存在时，并不要求判断哪个具体实例符合手术流程描述。如表1所示，现有手术器械数据集主要关注类别级分割，缺少对语言条件、结构化空间定位以及跨不同手术流程的实例级消歧的集成支持。相反，通用视觉-语言定位基准（例如 RefCOCO） [17, 18, 33] 无法体现手术环境固有的视觉复杂性、遮挡、细粒度器械形态以及领域特定约束。因此，手术场景中的定位式实例级推理仍基本未经测试。弥合这一差距需要一个基准，明确将自然语言指代解析与精确的像素级器械定位相结合。此类评估必须在真实的术中条件下进行，如图1所示。此类问题设定必须支持存在歧义的多器械场景，提供结构化空间定位标注以确保定位一致性，并采用能够量化定位准确度和分割精度的评估协议。

为满足这一需求，我们提出**GroundedSurg**，这是一个基于定位的手术器械分割基准，包含612张手术图像和1,071

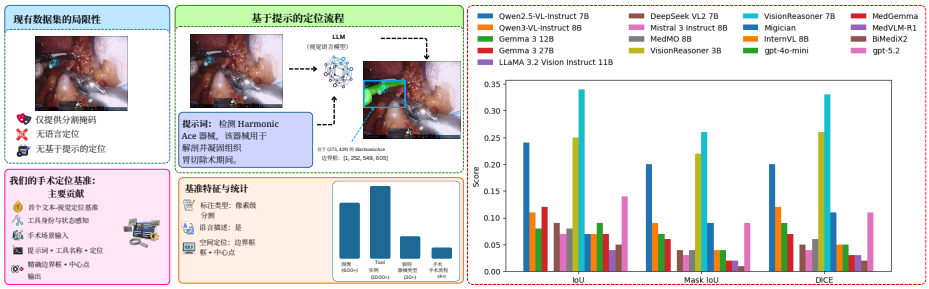

图1：**GroundedSurg**概览。(a) 现有数据集侧重于类别级分割，不具备语言条件或实例级定位。(b) GroundedSurg 引入了带有结构化空间标注的自然语言查询，用于查询条件下的器械定位。(c) 基线结果显示出显著的性能差距，凸显了基于定位的手术感知所面临的挑战。

项器械级标注，将手术器械感知重新定义为一项语言条件的实例级分割任务。给定一张手术图像和一条自然语言查询，该查询通过功能角色、空间关系或解剖结构交互来描述特定器械，目标是定位并分割满足该描述的器械实例。这一表述有别于传统类别级分割，它要求在视觉相似的器械之间进行显式消歧，并将上下文指代定位到单一的空间定位实例。GroundedSurg纳入了结构化空间标注，包括边界框和中心点，以在多个粒度层面量化定位准确度。边界框支持粗粒度实例验证，而像素级掩码则用于在遮挡、镜面反射、运动模糊和器械重叠等真实手术挑战下评估细粒度轮廓勾画。每个图像—查询对都被视为一个独立的评估单位，从而能够在多器械场景中进行受控的实例级指标计算。通过在标准化评估框架内统一自然语言指代、结构化空间定位和多术式多样性，GroundedSurg为基于定位的手术感知确立了一个严谨的基准，从而推动下一代具备定位感知能力的手术人工智能系统。

贡献。(1) 我们将手术器械感知重新定义为一项有定位能力的视觉—语言任务，要求解析依赖上下文的指代，以识别特定器械实例。**(2)** 我们提出GroundedSurg，这是一个手术基准，通过系统地将自然语言描述与明确的空间定位标注相结合，包括边界框、中心点和像素级掩码，从而能够严格评估语言条件下的实例级定位与分割。**(3)** 我们构建了一个多样化的多术式数据集，涵盖不同的手术领域和成像条件，为具备定位感知能力的术中视觉系统提供一个符合临床实际且可复现的测试平台。

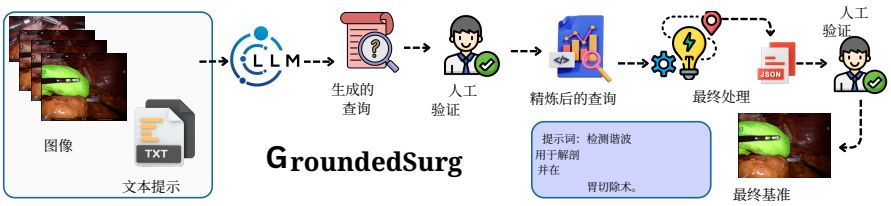

图 2：**GroundedSurg** 基准流水线。手术图像与初始提示词配对后，由视觉语言模型处理，以生成结构化的器械描述。所有**1071**条查询均经过人工和临床医生验证，以确保语义正确并消除歧义。最终标注以标准化 JSON模式存储，其中包含空间定位信息（边界框和中心点）以及分割掩码。

2 GroundedSurg

问题形式化：GroundedSurg 将手术器械感知形式化为一项语言条件下的实例级分割任务。基准中的每个实例均由一张手术图像及其对应的自然语言描述组成，该描述指向单个目标器械，并附有结构化空间定位标注和像素级分割掩码。

形式化地，设 $I \in \mathbb{R}^{H\times W\times 3}$表示一张高度为 H、宽度为 W的手术图像。对于每张图像，一个目标器械实例由以下内容定义：(i) 一条自然语言查询 T，通过器械的术中角色、空间关系或交互描述该器械；(ii) 一个边界框 $B = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$；(iii) 一个中心点 $C = (x_c, y_c)$；以及 (iv) 一个二值分割掩码 $M \in \{0,1\}^{H\times W}$。ˆ 目标是预测与 T所描述器械相对应的分割掩码 M，即学习映射 $f(I, T, B, C) \rightarrow M$。每个图像—查询对恰好对应一个器械实例。对于包含多个器械的场景，会针对每个指代分别构建实例，以确保监督信号明确无歧义，并支持受控的实例级评估。空间标注 (B, C) 作为辅助定位线索，可减少歧义，并加强语言指代与视觉定位之间的对齐。

基准设计：与传统的类别级分割不同，GroundedSurg 要求模型根据语言和空间线索识别并分割特定的器械实例；当多个外观相似的器械共存时，模型必须进行明确的消歧。边界框 B 提供粗定位监督，用于验证实例；中心点 C 则引入了额外的空间约束，以增强定位鲁棒性。尽管有这些锚点，主要目标仍是对所指器械进行精确的像素级勾画，以满足术中精度要求。将每个图像—查询对视为独立的评估单位，可避免不同实例之间的相互干扰，并实现规范的实例级评估。

数据集构建：该数据集汇集了来自公开手术数据集的样本，涵盖多种手术类型和成像条件，包括

表 2：GroundedSurg 数据集的汇总统计量。

统计量	数值
图像数量	∼612
标注数量	∼1,071
平均每张图像的器械数	∼1.6
外科手术	4
标注类型	像素级分割
空间定位	边界框 +中心点
语言描述	实例级指代

表 3：GroundedSurg 中使用的评估指标。

指标	描述
基于区域的指标	
IoU	$\frac{\lvert \hat{M} \cap M \rvert}{\lvert \hat{M} \cup M \rvert}$
IoU@0.5/0.9	交并比 $\geq$ 0.5/ 0.9
mIoU	平均交并比
Dice	$\frac{2\lvert \hat{M} \cap M \rvert}{\lvert \hat{M} \rvert + \lvert M \rvert}$
定位指标	
边界框交并比	边界框交并比（预测值与真实标注）
NDE	归一化中心距离

InSeg1 [6] 和 InSeg2 [6]（眼科），SISVE [32] （胃切除术），En- doVis [26]（机器人辅助肾切除术），以及 CholecInstanceSeg [1]（腹腔镜胆囊切除术）。这些数据来源涵盖了多种手术环境，从使用精细器械的显微外科场景，到具有关节式器械、杂乱背景和镜面反射的腹腔镜场景。这种多样性带来了真实的挑战，包括遮挡、器械重叠和较高的视觉相似度，从而能够在具有临床代表性的条件下进行评估。

标注协议： 每个基准实例对应一条器械标注，并配有经语言定位的自然语言提示和像素级分割监督。当同一图像中出现多个器械时，每个实例均单独标注。实例级监督。 每条标注都分配一个唯一标识符。主要监督信号是像素级掩膜，可精确勾勒目标器械的可见范围。所有掩膜均与原始图像分辨率对齐，并经过几何一致性验证。辅助空间定位信息（包括边界框和中心点标注）则提供结构化定位线索。 语言定位：查询由 Qwen-2.5 VL-Instruct [27] 根据视觉内容和空间标注生成。每条标注均包含：(i) 对目标器械的文字描述；(ii) 通过边界框和中心点参照进行明确的空间定位；以及 (iii) 用于说明分割目标的推理提示。

临床验证：该数据集遵循一套经过临床验证的半自动化流水线。临床医生会审查并核实由视觉语言模型生成的初始查询，以消除幻觉、纠正语义不一致，并确保其符合手术场景。精炼后的查询经过标准化后，整合到统一的 JSON模式中，随后再进行一轮人工验证，以确保提示与掩膜对齐准确且空间一致。

数据集统计： 该数据集包含约 **612 张手术图像** 和**1,071 条器械级标注**。由于每张图像中可能出现多件器械，标注数量多于图像数量。该数据集涵盖四种手术操作和多样的成像条件。所有标注均包含像素级掩码、边界框、中心点和自然语言描述，构成了一个中等规模、细粒度的基准，用于在真实临床变化条件下进行基于定位的手术分割。

表4：在统一的查询条件分割协议下，视觉语言模型在 GroundedSurg 上的定量比较。最佳结果以**粗体**显示，次佳结果带下划线。

模型	参数	IoU		mIoU@0.1					
开源模型									
Qwen2.5-VL-Instruct [27]	7B	**0.52**	**0.33**	0.24	**0.26**	**0.22**	0.20	1.45	0.20
Qwen3-VL-Instruct [3]	8B	0.36	0.13	0.11	0.17	0.12	0.09	1.78	0.12
Gemma 3 | [7] (12B)	12B	0.28	0.09	0.08	0.13	0.09	0.07	1.45	0.09
Gemma 3 [7] (27B)	27B	0.43	0.11	0.12	0.11	0.07	0.06	1.52	0.07
LLaMA 3.2 Vision [9]	11B	0.02	0.00	0.00	0.01	0.00	0.00	2.90	0.00
DeepSeek VL2 [31]	7B	0.28	0.10	0.09	0.06	0.05	0.04	1.89	0.05
Mistral 3 [13]	8B	0.27	0.06	0.07	0.07	0.04	0.03	1.89	0.04
推理导向模型									
VisionReasoner（3B） [14]	3B	0.20	0.06	0.25	0.23	0.03	0.22	1.39	0.26
VisionReasoner（7B） [14]	7B	0.32	0.20	**0.34**	0.25	0.12	**0.26**	**1.05**	**0.33**
Migician [12]	7B	0.19	0.08	0.07	0.16	0.12	0.09	1.72	0.11
InternVL [4]	8B	0.29	0.05	0.07	0.08	0.05	0.04	1.69	0.05
医学领域模型									
MedMO [5]	8B	0.33	0.04	0.08	0.08	0.05	0.04	1.50	0.06
MedGemma [24]	4B	0.29	0.04	0.07	0.05	0.03	0.02	1.88	0.03
MedVLM-R1 [20]	2B	0.13	0.01	0.04	0.05	0.03	0.02	2.16	0.03
BiMediX2 [19]	8B	0.23	0.04	0.05	0.07	0.02	0.01	2.11	0.02
闭源模型									
GPT-4o-mini [16]	–	0.35	0.04	0.09	0.08	0.05	0.04	1.63	0.05
GPT-5.2 [25]	–	0.39	0.20	0.14	0.16	0.12	0.09	1.95	0.11

评估协议：性能在实例级别进行评估，将每个图像—查询对独立处理。

基于区域的指标：使用交并比（IoU）评估分割质量，其定义为 IoU $= \frac{|M \cap M|}{|M \cup M|}$。我们报告 IoU@0.5、IoU@0.9、平均交并比（mIoU）和 Dice 系数，以评估重叠度和边界保真度。

定位指标：使用边界框交并比和预测中心点与真实中心点之间的归一化距离误差（NDE）评估空间定位准确率。所有指标均在实例层面取平均，以便系统地比较多器械手术场景中的语言条件分割模型。

3 实验

实验设置：我们采用统一的语言条件实例分割协议评估 GroundedSurg：模型预测结构化定位输出（边界框和中心点），并将其投射到冻结的基于 SAM 的后端，以获得最终掩码。所有模型均在不进行微调的零样本设置下评估，并针对每个图像—查询对报告实例级性能。

视觉语言模型间的比较： 表 4展示了在统一协议下，对开源、推理导向、医疗领域和闭源模型进行的定量比较。不同模型系列在更严格的重叠阈值下，性能仍然有限。虽然某些模型在 IoU@0.1 上取得了中等数值，但在

表5。以 SAM2 [21] 和 SAM3 [11]作为分割模型，并在 VLMS 检测结果基础上进行比较的代表性模型。最佳结果以**粗体**标出，次优结果以下划线标出。

模型	SAM2		SAM3	
	掩码交并比	Dice	掩码交并比	Dice
Qwen2.5	**0.22**	**0.26**	0.20	0.20
Qwen3	0.12	0.16	0.09	0.12
Gemma 3 27B	0.03	0.04	0.06	0.07
MedMO	0.00	0.01	0.04	0.06
gpt-5.2	0.04	0.06	0.09	0.11
VR-7B	0.17	0.19	**0.26**	**0.33**

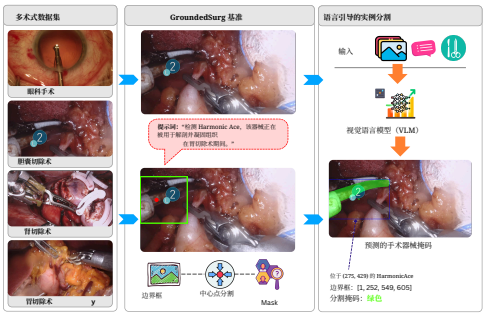

图3。GroundedSurg 的组成部分。

表5：两种提示词表述下的敏感性分析。每项指标的最佳结果以**粗体**显示，次优结果带下划线。

模型	提示1				Prompt2			
	IoU@0.1	IoU@0.3	边界框交并比	NDE ↓	IoU@0.1	IoU@0.3	边界框交并比	NDE ↓
Qwen2.5-VL-7B [27]	0.52	0.33	0.24	1.45	0.38	0.09	0.11	1.43
Qwen3-VL-8B [3]	0.36	0.13	0.11	1.78	0.11	0.02	0.09	2.94
Gemma3-27B [7]	0.43	0.11	0.12	1.52	0.42	0.08	0.10	1.51
InternVL-8B [4]	0.29	0.05	0.07	1.69	0.37	0.09	0.10	1.42
MedGemma-4B-it [24]	0.29	0.04	0.07	1.88	0.21	0.06	0.07	1.91
VisionReasoner-7B [14]	0.32	0.08	**0.34**	**1.05**	**0.58**	**0.39**	0.31	1.14

更高阈值（IoU@0.3 及以上）下，准确率会急剧下降，这表明模型有时能够实现粗定位，但精确的边界对齐仍然具有挑战性。

在开源模型中，Qwen2.5-VL [27]的粗粒度定位性能相对较强，但在更严格的重叠度标准下准确度有所下降，这表明其细粒度空间精度有限。相比之下，面向推理的模型表现出更好的定位一致性。VisionReasoner-7B [14] 取得了最高的边界框交并比和Dice分数，表明其空间定位能力和掩码保真度更强，也说明结构化推理增强了模型在手术场景歧义下的鲁棒性。医学领域模型并未始终优于通用模型，这表明仅靠领域预训练并不能保证实例级定位能力提升。闭源系统的表现虽具竞争力，但并非最佳，这进一步凸显了该基准的整体难度。

分割后端分析。为评估可提示分割后端的影响，我们在不同的冻结分割模型下比较了具有代表性的模型。表4显示，SAM2 [21] 与 SAM3 [11]之间的分割质量差异显著。虽然有些模型在 SAM3 下的掩码投影效果有所改善，但其他模型的提升微乎其微。值得注意的是，VisionReasoner 在 SAM3 下表现出显著提升，这表明精确的定位输出与先进的掩码解码具有更强的兼容性。这些结果凸显了定位准确度与分割投影在整个流水线中的紧密关联。

如表5所示，通用视觉语言模型在不同提示风格下的性能差异显著。例如，Qwen3-VL [3] 和其他几个模型在使用

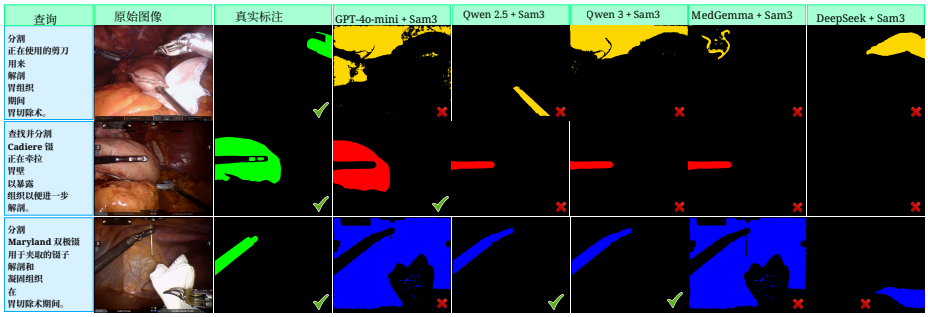

图 3：GroundedSurg 上的定性比较表明，在将结构化定位输出投射到冻结分割后端时，推理导向模型生成的掩码比通用模型在空间上更加精确，尤其是在多器械和视觉杂乱场景中。

备用提示词时，交并比和边界框交并比均显著下降。相比之下，推理导向模型表现出更强的鲁棒性；VisionReasoner-7B 在使用 Prompt2 时保持了性能，甚至有所提升，这表明其语义定位能力更强，对固定提示结构的依赖更低。这些发现表明，提示工程对于通用多模态模型仍至关重要，而以推理为核心的架构则对语言改述表现出更强的不变性。

定性分析。图3 展示了代表性模型之间的定性比较。通用视觉语言模型并不直接执行分割，因此常常在杂乱手术场景中生成粗略或不准确的空间定位。随后将这些预测区域提供给 SAM3 以生成掩码时，所得分割仍不精确，边界不准确且存在上下文泄漏——在多器械场景中尤为明显。这些定性观察与表4中报告的定量趋势一致。

提示词调优的影响。提示敏感性结果（表5）证实，分割性能会受到指令结构的显著影响。尽管许多模型在轻微改写时表现不稳定，推理导向模型仍能保持一致的定位行为，这表明显式空间推理能够提升模型对语言变异性的鲁棒性。

分割投影效应。SAM2与SAM3的比较（表4）进一步表明，掩码质量不仅取决于定位精度，还取决于后端投影特性。掩码解码的改进对定位能力更强的模型带来不成比例的收益，凸显了评估定位与分割组件的重要性。

4 结论

我们提出了 GroundedSurg，这是一个基于定位的基准，将手术器械感知重新定义为语言条件的实例级分割任务。通过整合自然语言指称与结构化空间

定位和像素级掩码，我们超越了类别级识别，转向在不同手术过程中实现依赖上下文且具有临床意义的定位。实验表明，当前多模态模型难以在复杂手术场景中实现可靠的实例级定位。虽然有时能够进行粗略定位，但精确的边界描绘以及对提示变化的鲁棒性仍然有限。GroundedSurg 提供了一个标准化且具有临床相关性的测试平台，可推动具备定位能力的术中视觉发展，并凸显了对能够更好地整合语言推理与细粒度空间感知的模型的需求。

参考文献

1. Alabi, O., Toe, K.K.Z., Zhou, Z., 等：CholecInstanceSeg：用于腹腔镜手术的器械实例分割数据集。Scientific Data**12**，825（2025年）。 https://doi.org/10.1038/s41597-025-05163-w 2. Allan, M., 等：2017年机器人器械分割挑战赛（2019年），https://arxiv.org/abs/190206426 3。. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., 等：Qwen3-vl技术报告。arXiv预印本 arXiv:2511.21631（2025年）4. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 等：InternVL：扩展视觉基础模型并使其适用于通用视觉语言任务。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第24185–24198页（2024年）5. Deria, A., 等：MedMO：面向医学图像的多模态大语言模型的定位与理解。arXiv预印本 arXiv:2602.06965（2026年）6. Fox, M., Taschwer, M., Schoeffmann, K.：使用Mask R-CNN对白内障手术视频进行基于像素的器械分割。载于：第33届IEEE计算机医学系统国际研讨会（CBMS 2020），美国明尼苏达州罗切斯特，2020年7月28日至30日。第565–568页。IEEE（2020年）。https://doi.org/10.1109/CBMS49503.2020.001127. Gemma团队：Gemma 3技术报告（2025年），https://arxiv.org/abs/2503.197868. Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., 等：CaDIS：用于图像分割的白内障数据集。arXiv预印本 arXiv:1906.11586（2019年），https://arxiv.org/abs/190611586 9。. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Let- man, A., Mathur, A., Schelten, A., Vaughan, A., 等：Llama 3模型家族。arXiv预印本 arXiv:2407.21783（2024年）10. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.：Cholec- seg8k：基于cholec80的腹腔镜胆囊切除术语义分割数据集12453 11202020123(),https://arxiv.org/abs/ 。. Jiang, C., 等：Medical SAM：用于提示驱动的通用医学图像分割的基础模型。arXiv预印本 arXiv:2601.10880（2026年），https://arxiv.org/abs/260110880 12。. Li, Y., Huang, H., Chen, C., Huang, K., Huang, C., Guo, 等：Migician：揭示多模态大语言模型中自由形式多图像定位的奥秘。载于：计算语言学协会论文集（ACL 2025）。第9845–9867页（2025年）

13. Liu, A.H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., 等：Ministral 3（2026年）14. Liu, Y., Qu, T., 等：VisionReasoner：通过强化学习实现统一的推理集成视觉感知。arXiv预印本 arXiv:2505.12081（2025年）15. Maier-Hein, L., 等：面向下一代干预的手术数据科学。《自然-生物医学工程》1，691–696（2017年）。https://doi.org/10.1038/s41551-017-0132-7 16. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.：生成和理解明确无歧义的物体描述。载于：IEEE计算机视觉与模式识别会议论文集。第11–20页（2016年）17. Mao, J., 等：生成和理解明确无歧义的物体描述，第11–20页（2016年）18. Mao, J., 等：生成和理解明确无歧义的物体描述。载于：IEEE计算机视觉与模式识别会议论文集。第11–20页（2016年）19. Mullappilly, S.S., 等：BiMediX2：适用于多种医学模态的生物医学专家大语言模型（2024 07769 20 2412 1），https://arxiv.org/ 。Pan, J., 等：MedVLM-R：通过强化学习激励视觉语言模型（VLM）的医学推理能力。arXiv预印本 arXiv:2502.19634（2025年）21. Ravi, N., 等：SAM 2：分割图像和视频中的任意对象（2024年），https: //arxiv.org/abs/2408.00714 22. Ross, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., 等：2019年鲁棒医学器械分割挑战赛（2020年），https: //arxiv.org/abs/2003.10299 23. Sachdeva, B., Akash, N., Ashraf, T., Mueller, S., Schultz, T., Wintergerst, M.W.M., Prasad, N.S., Murali, K., Jain, M.：面向手动小切口白内障手术的阶段感知工具分割（2024 16794 24 2411 2507），https://arxiv.org/ 。Sellergren, A., 等：MedGemma技术报告。arXiv预印本 arXiv: .05201（2025年）25. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., 等：OpenAI GPT-5系统卡。arXiv预印本 arXiv:2601.03267（2025年）26. Sun, L., Chen, X.：利用数据集级样本队列，在内窥镜机器人手术视频中通过像素级对比学习进行多类别器械分割。《IEEE Access》12，156867–156877（2024年）27. Team, Q.：Qwen2.5：一系列基础模型（2024年9月），https://qwenlm.github.io/blog/qwen2.5/ 28. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.：EndoNet：用于腹腔镜视频识别任务的深度架构（2016 03012 29 1602），https://arxiv.org/ 。Wang, Q., Zhao, S., Xu, Z., Zhou, S.K.：LACOSTE：利用立体与时间上下文进行手术器械分割。《医学图像分析》99，103387（11 2024 103387 30 10 1016 2024）。https://doi.org/ 。/j.media. 。Wang, Z., 等：AutoLaparo：用于腹腔镜子宫切除术图像引导手术自动化的集成多任务新数据集（2022年），https://arxiv. org/abs/2208.02049 31. Wu, Z., 等：DeepSeek-VL2：面向高级多模态理解的混合专家视觉语言模型（2024年），https://arxiv.org/abs/2412.10302

32. Yoon, J., 等：使用语义图像合成和虚拟手术环境进行手术场景分割：通过对象尺寸感知随机裁剪和背景标签增强提升照片级真实感合成（2024年3月），https://sisvse.github.io/ 33. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.：指代表达式中的上下文建模。CoRR **abs/1608.00272**（2016年），http://arxiv.org/abs/1608.00272