SurgBench：用于手术视频分析的统一大型基准测试

Jianhui Wei[1,4,*] **Zikai Xiao**[1,4,*] **Danyu Sun**[1] **Luqi Gong**[2]
Zongxin Yang[3] **Zuozhu Liu**[1,4,†] **Jian Wu**[1,4,†]
[1]Zhejiang University [2]Zhejiang Lab [3]Harvard University
[4]Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence
[*]jianhui1.24@intl.zju.edu.cn [†]zuozhuliu@intl.zju.edu.cn

Abstract

外科手术视频理解对于实现自动化术中决策、技能评估以及术后质量改进至关重要。然而，外科手术视频基础模型的发展仍受限于用于预训练和系统评估的大规模、多样化数据集的匮乏。在这篇论文中，我们介绍了SurgBench，这是一个统一的外科手术视频基准框架，包括一个预训练数据集 SurgBench-P 和一个评估基准 SurgBench-E。SurgBench 提供了广泛覆盖多样化的外科手术场景，SurgBench-P 涵盖了 53 百万帧、涉及 22 种外科手术和 11 个专业领域，而 SurgBench-E 则提供针对六个类别（阶段分类、相机运动、工具识别、疾病诊断、动作分类和器官检测）以及 72 个细粒度任务的稳健评估。大量实验表明，现有的视频基础模型在各种外科手术视频分析任务中难以泛化，而在 SurgBench-P 上进行预训练则显著提升性能，并在未知手术和模式上表现出优越的跨域泛化。我们的数据集和代码可根据要求提供。

1 介绍

外科手术视频分析正迅速成为现代外科护理进步的基石，提供了对手术过程复杂性的洞察 (Green et al., 2019) 。系统地解读这些视频的能力可以显著影响术中决策、自动化外科技巧评估，并为术后质量改进计划提供信息 (Loftus et al., 2020; Prebay et al., 2019; Grenda et al., 2016; Levin et al., 2019; Grüter et al., 2023; Dimick and Varban, 2015) 。

为了应对临床机遇，视频基础模型（VFMs）作为外科视频分析的一个强大框架应运而生。通过利用大规模、多样化的数据集，VFMs 实现了高效且准确的视频建模，增强了外科视频分析的潜力并改善了外科实践 Wang et al. (2024); Zhao et al. (2024); Li et al. (2023); Madan et al. (2024) 。VFMs 在各种下游任务中展示了成功的表现，这得益于对包含数百万视频片段和数 TB 数据的庞大数据集的预训练 (Wang et al., 2024; Zhang et al., 2025; Tong et al., 2022; Wang et al., 2023a) 。这些数据集涵盖了多种视频类别，如 YouTube 视频、电影预告片和监控录像，使模型能够在不同领域中开发出可泛化的表示 (Li et al., 2023; Madan et al., 2024) 。

尽管 VFMs 在一般视频分析中已被证明有效，但它们在外科视频分析中的应用仍处于初期阶段，主要由于现有预训练和评估数据集中疾病类型、外科手术和专科的多样性有限，以及任务的全面性不足。疾病的多样性受到限制，因为大多数数据集关注特定病症，如结直肠癌 (Misawa et al., 2021) ，这限制了模型在不同解剖变异、合并症和罕见病中的泛化能力 (Bar et al., 2020) 。手术过程或专科的覆盖范围很窄，数据集主要表现特定的专科和微创手术，如腹腔镜胆囊切除术 Wang et al. (2022); Twinanda et al. (2016) 或机器人前列腺切除术 Ahmidi et al. (2017) ，而对开放式、混合式或不太标准化的手术技术代表不足。任务全面性的不足显而易见，因为倾向于强调孤立的分析任务，如特定手术过程的阶段分类 (Goodman

[*]Co-first author.
[†]Co-corresponding author.

Preprint.

Datasets	Pretraining Data			Evaluation Data
	# Surgical specialties	# Surgical procedures	# Frames	
Endo-FM (Wang et al. (2023b))	3	3	5M	×
GSViT (Schmidgall et al. (2024))	3	28	70M	×
Surg-3M (Che et al. (2025))	5	35	3M	×
SurgBench (Ours)	11	22	53M	✓ (72 Tasks)

Table 1: 用于手术视频预训练和评估的数据集比较。我们的基准在覆盖手术专科和手术程序方面具有优势，同时也为下游任务提供了全面的评估标准。

et al., 2024; Fujii et al., 2024) 或器械识别 (Ma et al., 2021)，而忽视了跨越多个时间和语义维度的综合临床工作流程。

在本文中，我们介绍了 SurgBench，这是一个统一的大规模外科视频基准框架，由一个预训练数据集 SurgBench-P 和一个评估基准 SurgBench-E 组成。预训练子集 SurgBench-P 由来自 16 个不同来源的 5300 万帧组成（如图 **??** 所示），涵盖了腹腔镜、内窥镜、机器人和开放手术 4 种主要外科手术方式。它跨越了 22 种不同的手术操作（例如阑尾切除术、胆囊切除术、结肠切除术等），涉及 11 个医学专科（见表 6 ），直接缓解了操作和疾病类型的同质性问题。我们进一步利用统一的时空协议（例如长度、帧率、编码），与诸如 Kinetics 等一般视频理解基准对齐，适用于所有视频来源。这种标准化确保了预训练的一致性，并支持从一般领域视频基础模型的有效知识转移。与预训练语料库互为补充，SurgBench-E 作为一个综合的微调和层次评估框架，专门设计以丰富任务多样性并严格评估临床实用性。SurgBench-E 包含 23004 个手术视频剪辑，其中有 6 个不同外科理解任务类别、10 个亚类的详细注释，并进一步细分为 72 项任务，为全面的模型评估提供了结构化的手段，详细内容见表 3 。

初步的实证验证强调了这一方法的有效性：在 SurgBench 上预训练的 SurgMAE，在后续任务中相比于在自然视频数据集 Kinetics 上训练的模型，表现出 7.9 % 的性能提升。我们的主要贡献如下：

- 我们引入了 SurgBench-P，一个大规模且标准化的预训练数据集（5300 万帧，22 个手术过程，11 个专科，4 种模式），旨在解决手术视频分析中过程和疾病类型的同质性问题。
- 我们构建了 SurgBench-E，这是一个涵盖 72 个细粒度任务的综合评估基准，分为 6 个类别，旨在减少任务碎片化并促进临床应用的系统评估。
- 我们使用 VideoMAE 验证了一种自监督预训练流程，以获得卓越的性能，展示了 SurgBench 在推进外科研究和教育方面的潜力。

2 相关工作

外科视频分析。早期的外科视频分析研究主要集中在单一疾病或有限的任务范围，例如外科阶段识别（Yu et al. (2018) ）、器械检测（Yu et al. (2018) ）和息肉检测（Ma et al. (2021) ）。这些方法通常依赖于小规模、单中心的数据集，缺乏多样性，导致其模型训练依赖于特定任务和情景，通用能力有限。例如，已经针对特定疾病开展内镜手术的研究（Nwoye et al. (2022) ）、腹腔镜子宫切除手术（Wang et al. (2022) ）、用于息肉检测的结肠镜视频（Mesejo et al. (2016) ，Ma et al. (2021) ）以及使用 PillCAM 数据的胶囊内镜检查（Smedsrud et al. (2021) ）。此外，由于标注成本高，许多这些研究工作都是高度监督的，因此难以扩展到更广泛的临床环境或多任务学习场景。

外科基础模型。随着大规模外科视频数据的不断增多，研究人员开发了外科基础模型 (Wang et al., 2023b; Schmidgall et al., 2024) ，通常利用自监督学习在多样化的数据集上进行，以获得可泛化的表示。这些模型表现出卓越的稳健性和跨域可转移性。最近，结合视觉语言和具身 AI 方法 (Li et al., 2024; Wang et al., 2025; Li et al., 2025c,a,b; Bi et al., 2024, 2025; Zeng et al., 2024; Guo et al., 2025; Ma et al., 2024; Zhou et al., 2025a,b) 进一步增强了复杂外科任务的语义理解和可解释性。

在本节中，我们介绍了构建 SurgBench 的方法，它包含用于预训练模型的 SurgBench-P 和用于微调及评估模型的 SurgBench-E。整个构建流程如图 **??** 所示。

Source ID	Source Name	Disease Type	Procedures	Task Type	Pre-train Frames	Evaluation Frames
S1	AVOS (Goodman et al. (2024))	Multiple	Multiple, Open	Phase/Action/Inst cls.	28,473,879	–
S2	AIxSuture (Hoffmann et al. (2024))	N/A (Skills)	Tissue Suturing	GRS assessment	2,867,078	–
S3	Cholec80 (Yu et al. (2018))	Gallbladder	Cholecystectomy	Phase cls., Tool presence	4,612,530	–
S4	CholecT45 (Nwoye et al. (2022))	Gallbladder	Cholecystectomy	Action cls. (triplets)	74,855	–
S5	Colonoscopic	Colorectal lesions	GI diagnosis	Disease cls.	36,534	–
S6	Endo-FM (Wang et al. (2023b))	Various GI	Various Endoscopy	SSL tasks	3,646,432	–
S7	SimSurgSkill 2021	N/A (Simulation)	Liver/Abdominal (sim)	Skill metrics, Tool detection	1,248,156	–
S8	JIGSAWS (2017) (Ahmidi et al. (2017))	N/A (Skills)	Suture/Knot/Needle	Skill assessment, Gesture classification	569,048	537,645
S9	CholecT50 (Nwoye et al. (2022))	Gallbladder	Cholecystectomy	Action cls., Phase cls., Inst detection	90,444	207,169
S10	AutoLaparo (Wang et al. (2022))	Uterine	Hysterectomy	Phase cls., Motion prediction, Segmentation	2,155,843	160,221
S11	EndoVis 2019 (Wagner et al. (2021))	Gallbladder	Cholecystectomy	Phase/Action/Instrument cls., Skill assessment	4,501,791	1,176,000
S12	Hyper-Kvasir (Borgli et al. (2020))	GI pathologies	GI procedures	Tissue/Pathology segmentation, Classification	889,372	102,515
S13	Colonoscopic-web (Mesejo et al. (2016))	Colorectal lesions	GI diagnosis	Data augmentation	75,298	75,298
S14	Kvasir-Capsule (Smedsrud et al. (2021))	GI pathologies	GI screening	Pathology clas.	4,765,114	61,760
S15	LDPolypVideo (Ma et al. (2021))	Colonic polyps	Polyp screening	Polyp detection, Classification	878,487	543,777
S16	Private Data	Right colon ca.	Lap. Rt. Hemicolectomy	Phase cls., Skill/Quality assessment	1,138,833	816,952
Total					56,062,458	3,680,417

Table 2: SurgBench 数据集的原始来源摘要，包括公开可用的学术数据集、医学竞赛数据集、演示数据和私有数据。来源涵盖不同的疾病类型、程序和任务类型。虚线以上的来源不提供标签，仅用于创建 SurgBench-P。

GI: Gastrointestinal; GRS: Global Rating Score; cls.: classification; sim: simulation; SSL: Self-supervised learning; Obj.: Object; ca.: cancer; Lap.: Laparoscopic; Rt.: Right; Inst.:Instrument

我们获取了大多数可公开获得的手术视频分析数据集，包括 15 个公共数据集和 1 个额外的私人数据集，扩展到 22 种手术过程、11 个手术专科以及相关任务，以构建 SurgBench。完整的数据集描述如表 2 所示。详细的手术专科和手术过程如表 6 所示。相比于现有的 工作，经常集中于特定的手术类型或疾病，我们将所有这些数据集统一在一起，旨在建立一个综合的手术视频分析数据集，以获得更具普遍性的模型性能。然而，SurgBench 中的某些数据集在发行时附带许可证，不允许进行二次分发。特别是，AVOS、SimSurgSkill2021 和 AIxSuture 使用更严格的许可证（例如，CC-BY-NC-ND 4.0），禁止修改和再分发。虽然这些数据集被用于模型的预训练，但它们未包含在我们作为基准发布的 SurgBench-E 中。希望使用这些数据集进行模型预训练的研究人员应直接从数据集提供者处申请访问。伦理考虑和原始数据集许可证在附录 A 中充分阐述。

2.1 数据集预处理

不同的外科手术视频在编码格式、FPS、视频时长、空间分辨率等方面存在差异。为了标准化数据管理并加快训练过程，我们在 16 个视频源中执行了一系列预处理和标准化步骤。我们使用 H.264 编码标准化视频，该编码与各种处理库提供了最佳兼容性，显著加快了训练管道中的数据加载。分辨率被压缩到最小尺寸为 320（保持纵横比，为数据增强提供空间，网络输入分辨率设定为 224x224），并且帧速率统一为 20-30 FPS。

对于格式标准化的视频，预训练和评估视频的处理方式有所不同。对于预训练数据，我们采用了一种相当简单的方法来处理预训练数据，通过将未标记的视频分割成持续时间为 10 秒的短片。针对评估数据，我们实施了一系列措施以确保基准剪辑的合理性和可用性。首先，我们将所有视频分为带标签的短片（持续 1-10 秒）。然后，将它们按 5:5 的比例分为训练集和测试集。为了防止模型关注与标签无关的背景或场景特定属性，来自同一视频的剪辑只能出现在训练集中或测试集中。为了防止标签过于主导或稀少污染基准，IF（不平衡因子）通过动态控制剪辑持续时间、下采样主导标签的样本并去除稀少标签来保持在 10 以内。对于一个视频对应多个标签的情况，我们将多标签样本分割成多个单标签样本，并使用 top-k 准

确度指标。我们通过在这些数据集上微调 VideoMAE（标准版）测试了每个数据集的效能，训练动态如图 **??** 所示。训练和测试视频的标签分布相对接近，呈现出长尾分布，如图 1 所示。各数据集处理步骤的详细信息见附录 C 。

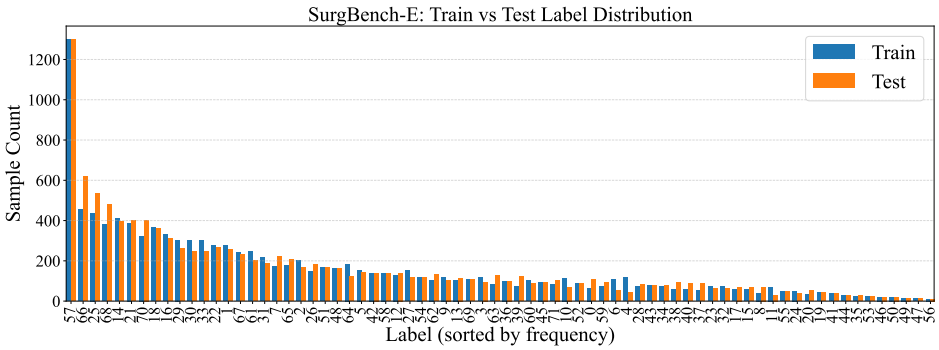


Figure 1: SurgBench-E 的训练和测试标签分布。明显的长尾模式与真实临床场景一致，既体现真实性又体现挑战性。标签对应的任务描述如表 7 所示。

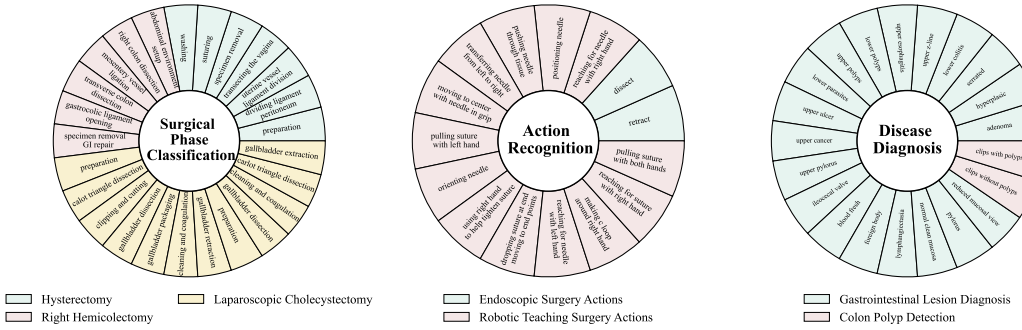
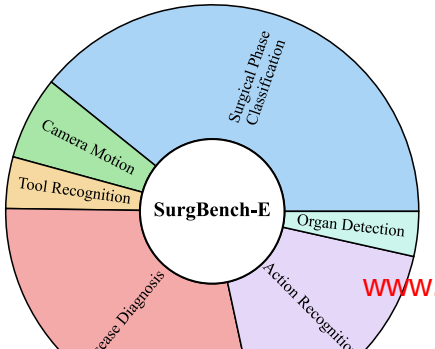


Figure 2: SurgBench-E 中任务的饼状图。我们有六个类别，这里显示了三个类别分布。不同的颜色代表不同的子类别。其他三个类别在附录的图 6 中提供。

整个 SurgBench 包含 225,250 个视频片段，59,742,875 帧，总时长 575 小时，涵盖了 22 种外科手术和 11 个专业。SurgBench-E 总共包含 23,004 个视频片段和 72 个细粒度标签，用于便于微调和测试模型在广泛外科任务上的能力。我们在 SurgBench-E 上定义了一个层次分类法，称为 6C-10S-72T，包含 6 个类别（其中三个在图 3 , 图 2 中示例）和 10 个子类别，以及 72 个详细任务，如表 3 所列。10 个子类别的样本分布自然呈长尾分布，如图 **??** 所示。样本大小与每个类别中的标签数正相关。

3 方法论

持续预训练我们通过利用在通用领域 Kinetics-400 数据集上预训练的 VideoMAE 模型开始我们的持续预训练（CPT）。用于预训练的数据 SurgBench-P（总共 74.4 百万帧）经过了四阶段的精炼，以确保它能够从大规模数据中学习通用表示，同时与下游任务保持一致。训练步骤包括：（1）最初收集所有可用视频至 225,250 个剪辑；（2）过滤掉不太相关或过于占优势的大规模样本（例如来自 AVOS 的）；（3）对代表性不足的数据应用上采样以鼓励更多的 IID 学习；（4）进行最终精确的以 IID 为导向的阶段，同时进行上采样和下采样。最终获得了一组精炼的 39,807 个视频剪辑（调整为 224 × 224 分辨率）用于最终的 CPT 阶段。遵循 VideoMAE 方法，我们采用了一个不对称的编码器-解码器架构。预训练任务涉及使用极高的掩码比率 0.9 重构随机掩码的空间时间“管”片段。对两种 VideoMAE 变体执行了 CPT：VideoMAE-Standard，其包含一个 ViT-Base 编码器和一个 4 层 Transformer 解码器，以及 VideoMAE-Large，其特点是



Category	Sub-Category (SC)	# Task	SC Code	Source ID
Phase Classification	Hysterectomy	7	T1	S10
	Right Hemicolectomy	6	T2	S16
	Laparoscopic Cholecystectomy	12	T3	S9, S11
Camera Motion	Camera Motion	7	T4	S10
Tool Recognition	Tool Recognition	2	T5	S9
Disease Diagnosis	Gastrointestinal Lesion Diagnosis	19	T6	S12, S13, S14
	Colon Polyp Detection	2	T7	S15
Action Classification	Endoscopic Surgery Actions	2	T8	S9
	Robotic Teaching Surgery Actions	13	T9	S8
Organ Detection	Organ Detection	2	T10	S9

Table 3: 我们的 SurgBench-E 的 6C-10S-72T 分类。Source ID 指的是本任务中所用数据集的来源。表格中使用的 SC 代码见 4 。

配备 ViT-Large 编码器和 12 层 Transformer 解码器。两种变体的 CPT 参数包括每个剪辑 16 帧的输入，时间采样率为 4。我们使用了 AdamW 优化器（$\beta_1 = 0.9, \beta_2 = 0.95$）。Base 模型的有效批量大小为 512（每 GPU 批量 64），Large 模型的有效批量大小为 256（每 GPU 批量 32）。结果模型被命名为 SurgMAE-CPT（标准版和大版）。

微调：我们在 SurgBench-E 上微调和测试 SurgMAE-CPT，所得模型称为 SurgMAE（标准和大型）。为验证持续预训练在骨干特征提取器上的有效性，我们选择训练分类器并分析训练动态。我们使用 AdamW 优化器（$\beta_1 = 0.9$，$\beta_2 = 0.99$），学习率为 1e-3，批量大小为 64 来微调标准 SurgMAE-CPT，而大型 SurgMAE-CPT 采用的学习率为 2e-3，批量大小为 16。

训练成本：对于持续预训练，实验在一个由 8 个 NVIDIA A100 GPU 组成的集群上进行。VideoMAE-Base 模型在手术数据上持续预训练了总共 54 个周期，而 VideoMAE-Large 模型训练了 38 个周期。两种版本的 CPT 过程共花费了一周时间。对于微调，我们使用了一台拥有 8 个 RTX3090（24 GB）GPU 的主机。标准和大模型都训练了 100 个周期，收敛耗时 14 小时。

4 实验与结果

在本节中，我们展示了实验设置和结果，以验证 SurgBench 在外科视频分析中的实用性。我们评估了使用自监督预训练的 VideoMAE 在 SurgBench-E 上微调的模型性能。我们的实验集中在理解训练动态、模型可扩展性、超参数的影响、泛化能力以及持续预训练的好处。

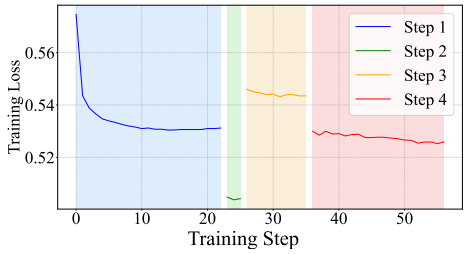
4.1 主要实验结果

我们使用 VideoMAE 架构在 SurgBench-P 数据集上进行了预训练（所有参数都进行了微调），并进一步在 SurgBench-E 上进行了微调。在表格 4 中总结了 10 个子类别的性能。为了减轻单个样本中存在多种类别可能导致结果不稳定的干扰，我们考虑了 top-1 准确率和 top-3 准确率。为了比较不同骨干网络的表示能力，我们在微调时冻结了特征提取器，只训练了分类头。

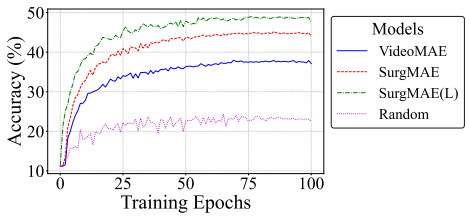
结果表明，在手术视频上持续进行预训练显著提高了 SurgBench-E 的准确性，各项指标均显示出持续增长。具体来说，top-1 准确性提高了 7 %，top-3 准确性提高了 7.9 %。除了 T4 和 T8 之外，大多数其他子类别也显示出明显的改进，这两个类别的性能几乎没有变化。这表明预训练数据具有可靠的质量，并且自监督预训练有助于学习稳健的表示。

Task	Random Init.		VideoMAE		SurgMAE		SurgMAE(L)	
	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3	Top-1	Top-3
T1	0.000	0.162	0.125	0.356	0.185	0.484	0.182	0.501
T2	0.364	0.667	0.423	0.777	0.440	0.818	0.493	0.856
T3	0.126	0.523	0.379	0.653	0.399	0.671	0.422	0.742
T4	0.000	0.000	0.298	0.629	0.293	0.628	0.306	0.645
T5	0.395	0.799	0.287	0.766	0.319	0.749	0.324	0.760
T6	0.413	0.605	0.569	0.744	0.645	0.838	0.666	0.905
T7	0.003	0.045	0.374	0.534	0.561	0.791	0.739	0.930
T8	0.000	0.044	0.005	0.254	0.015	0.383	0.020	0.411
T9	0.003	0.408	0.102	0.446	0.475	0.743	0.627	0.825
T10	0.000	0.290	0.122	0.632	0.200	0.575	0.279	0.695
Avg.	0.229	0.478	0.378	0.652	0.448	0.731	0.487	0.788

Table 4: 模型在不同任务（T1-T10）上的 Top-1 和 Top-3 准确率。粗体表示每个任务的最佳性能。除（L）外，模型均为标准大小，（L）代表大型。



(a) 在 SurgBench-P 上，标准 SurgMAE 的连续预训练动态跨越 4 个步骤。



(b) 在 SurgBench-E 上微调不同模型的动态。该模型为标准尺寸，除了 (L)，代表大尺寸版本。

Figure 4: 在 SurgBench 上预训练和微调的动态。对于预训练，我们跟踪四个步骤，从大规模数据到小规模分布对齐，使用 SurgBench-E。对于微调，我们比较随机初始化的权重、在 Kinetics 上预训练的 SurMAE，以及在 SurgBench-P 上预训练的大型/标准版本。损失曲线的稳定下降和准确率的提高证明了我们的数据集的合理性和质量。

为了详细了解我们的预训练和小任务集上具体的类别表现波动，我们绘制了每个类别的性能指标，如表 **??** 所示。我们发现了两个关键模式。首先，某些类别的表现显著低。这归因于某些稀有类型在一些视频领域的稀缺，导致数据量极为有限。这为未来的模型优化带来了挑战：如何在狭义定义或代表性不足的类别中实现高性能。其次，预训练过程持续提高了几乎所有单个类别的表现，显示出模型强大的泛化能力。

4.2 训练动态

一个稳定的训练损失曲线可以强烈地暗示数据的质量和训练方案的有效性。在预训练期间，为了评估数据和训练策略的有效性，我们分析了 VideoMAE 标准版本中四个渐进数据分布阶段的损失动态。我们发现，在每个阶段内，损失都非常平稳地下降，如图 4 所示。此外，由于训练数据的难度和分布的差异，不同阶段之间损失值的过渡并不平滑。然而，在实际任务测试中，这种不平滑并不会对性能产生负面影响。

此外，我们观察到，即使在最后阶段损失曲线并没有显著下降，它仍然有助于逐步稳定并改善任务微调的性能。这表明，即使损失下降较慢，模型仍然可以继续进行表示学习。

同样地，我们也通过训练动态检查了下游任务标签的质量。我们发现，随着训练周期数量的增加，准确率稳步上升。这一现象在标准模型、大参数模型、Kinetics 预训练模型和随机初始化模型中是一致的。

在语言预训练模型中，具有更多参数和更大规模的模型通常表现出更好的性能和更强的泛化能力。在外科基础模型的背景下，我们也测试了标准尺寸和更大模型在下游任务中的表

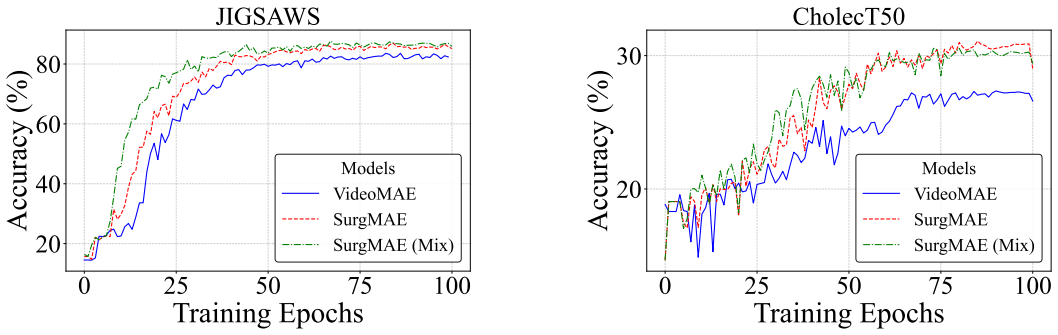
现。如表 4 所示，SurgMAE 和 SurgMAE(L) 在 top-1 准确率上提高了 3.9 %，在 top-3 准确率上提高了 5.7 %。

此外，在任务维度上观察到了持续的增长。这一现象与大规模语言模型的表现一致。此外，关于手术程序理解的后续研究表明，进一步增加模型规模可以导致更高的性能。

4.3 Surg-FM 能够泛化到未见过的案例吗？

在 SurgBench 上训练的模型可以改善在不同分布上的泛化能力。我们评估了在与 SurgBench 相似度不同的数据集上的表现。具体而言，我们在 LapGyn4 上进行测试，这是一个来自妇科腹腔镜手术的综合数据集，分类为四个不同的任务（手术操作、解剖结构、解剖操作和器械数量）(Leibetseder et al., 2018) 。我们专注于如 (Nasirihaghighi et al., 2024) 中所述的腹腔镜妇科事件识别任务。微调动态如图 **??** 中所示。值得注意的是，即便是在这种不相似的分布上，预训练在 SurgBench 上的模型依然展示了显著的性能提升，达到了 2.69 %。

在训练过程中，数据混合显著影响下游任务的表现。在 SurgBench-P 中，我们将所有源数据结合用于预训练，并采用了一种从“大规模 + 不受限分布”到“小规模 + 对齐分布”的过渡策略。我们假设，只要训练分布在最终阶段与下游任务分布对齐，就能获得最佳性能。使用数据集（JIGSAWS 和 CholecT50）继续在 Kinetics-400 上进行预训练和微调，我们观察到我们的模型在收敛速度和最终性能上达到了优越的水平，如图 5 所示。



(a) 对基于混合数据预训练的 SurgMAE 进行微调。我们首先使用 JIGSAWS 和 CholecT50 的组合数据集预训练模型，然后仅在 JIGSAWS 上进行微调。

(b) 在 CholecT50 上的 SurgMAE 微调动态。在结合 JIGSAWS 和 CholecT50 数据进行预训练后，该模型专门在 CholecT50 数据集上进行了微调。

Figure 5: 预训练于混合外科数据的模型与从头开始训练的模型之间微调性能的比较，展示了我们的数据混合策略的有效性。

5 结论

我们提出了 SurgBench，这是一个用于外科手术视频理解的综合基准。我们的贡献包括：(1) SurgBench-P，一个包含 22 种手术程序和 11 个外科领域的 5300 万帧的多样化预训练数据集；(2) SurgBench-E，一个结构化的评估基准，包含 6 个类别、10 个子类别和 72 个细粒度任务，促进全面的外科手术视频基准测试；以及 (3) 通过自监督学习在外科手术视频上显著提高性能的实证验证。实验结果表明，在 SurgBench 上预训练的模型优于在自然视频数据集上预训练的模型。SurgBench 提供了一个统一的平台用于外科手术视频分析，将加速该领域的研究进展。

尽管 SurgBench 作出了贡献，但它存在几个局限：(1) 长尾类分布对模型训练构成挑战，需要进一步研究以提高少数类的表现；(2) 语言监督整合尚未被深入探索，但可能具有益处；(3) 手术视频基础模型的最佳结构设计需要进一步研究，以最大化跨任务性能。

References

Ahmidi, N., Tao, L., Sefati, S., Gao, Y., Lea, C., Haro, B. B., Zappella, L., Khudanpur, S., Vidal, R., and Hager, G. D. (2017). A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041.

Bar, O., Neimark, D., Zohar, M., Hager, G. D., Girshick, R., Fried, G. M., Wolf, T., and Asselmann, D. (2020). Impact of data on generalization of ai for surgical intelligence applications. *Scientific reports*, 10(1):22208.

Bi, J., Wang, Y., Chen, H., Xiao, X., Hecker, A., Tresp, V., and Ma, Y. (2024). Visual instruction tuning with 500x fewer parameters through modality linear representation-steering. *arXiv preprint arXiv:2412.12359*.

Bi, J., Wang, Y., Yan, D., Xiao, X., Hecker, A., Tresp, V., and Ma, Y. (2025). Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*.

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., Jha, D., Eskeland, S. L., Randel, K. R., Pogorelov, K., Lux, M., Nguyen, D. T. D., Johansen, D., Griwodz, C., Stensland, H. K., Garcia-Ceja, E., Schmidt, P. T., Hammer, H. L., Riegler, M. A., Halvorsen, P., and de Lange, T. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283.

Che, C., Wang, C., Vercauteren, T., Tsoka, S., and Garcia-Peraza-Herrera, L. C. (2025). Surg-3m: A dataset and foundation model for perception in surgical settings. *arXiv preprint arXiv:2503.19740*.

Dimick, J. B. and Varban, O. A. (2015). Surgical video analysis: an emerging tool for improving surgeon performance.

Fujii, R., Hatano, M., Saito, H., and Kajita, H. (2024). Egosurgery-phase: a dataset of surgical phase recognition from egocentric open surgery videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 187–196. Springer.

Goodman, E. D., Patel, K. K., Zhang, Y., Locke, W., Kennedy, C. J., Mehrotra, R., Ren, S., Guan, M., Zohar, O., Downing, M., et al. (2024). Analyzing surgical technique in diverse open surgical videos with multitask machine learning. *JAMA surgery*, 159(2):185–192.

Green, J. L., Suresh, V., Bittar, P., Ledbetter, L., Mithani, S. K., and Allori, A. (2019). The utilization of video technology in surgical education: a systematic review. *journal of surgical research*, 235:171–180.

Grenda, T. R., Pradarelli, J. C., and Dimick, J. B. (2016). Using surgical video to improve technique and skill. *Annals of surgery*, 264(1):32–33.

Grüter, A. A., Van Lieshout, A. S., van Oostendorp, S. E., Henckens, S. P., Ket, J. C., Gisbertz, S. S., Toorenvliet, B. R., Tanis, P. J., Bonjer, H. J., and Tuynman, J. B. (2023). Video-based tools for surgical quality assessment of technical skills in laparoscopic procedures: a systematic review. *Surgical endoscopy*, 37(6):4279–4297.

Guo, H., Ma, Z., Zeng, Z., Luo, M., Zeng, W., Tang, J., and Zhao, X. (2025). Each fake news is fake in its own way: An attribution multi-granularity benchmark for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 228–236.

Hoffmann, H., Funke, I., Peters, P., Venkatesh, D. K., Egger, J., Rivoir, D., Röhrig, R., Hölzle, F., Bodenstedt, S., Willemer, M.-C., et al. (2024). Aixsuture: vision-based assessment of open suturing skills. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):1045–1052.

Leibetseder, A., Petscharnig, S., Primus, M. J., Kletz, S., Münzer, B., Schoeffmann, K., and Keckstein, J. (2018). Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology. In *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018*, pages 357–362. ACM.

Levin, M., McKechnie, T., Khalid, S., Grantcharov, T. P., and Goldenberg, M. (2019). Automated methods of technical skill assessment in surgery: a systematic review. *Journal of surgical education*, 76(6):1629–1639.

Li, J., Skinner, G., Yang, G., Quaranto, B. R., Schwaitzberg, S. D., Kim, P. C., and Xiong, J. (2024). Llava-surg: towards multimodal surgical assistant via structured surgical video learning. *arXiv preprint arXiv:2408.07981*.

Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., and Qiao, Y. (2023). Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960.

Li, Y., He, H., Cao, Y., Cheng, Q., Fu, X., and Tang, R. (2025a). M2IV: Towards efficient and fine-grained multimodal in-context learning in large vision-language models.

Li, Y., Yang, J., Li, B., and Tang, R. (2025b). CAMA: Enhancing multimodal in-context learning with context-aware modulated attention.

Li, Y., Yun, T., Yang, J., Feng, P., Huang, J., and Tang, R. (2025c). TACO: Enhancing multimodal in-context learning via task mapping-guided sequence configuration.

Loftus, T. J., Filiberto, A. C., Li, Y., Balch, J., Cook, A. C., Tighe, P. J., Efron, P. A., Upchurch Jr, G. R., Rashidi, P., Li, X., et al. (2020). Decision analysis and reinforcement learning in surgical decision-making. *Surgery*, 168(2):253–266.

Ma, Y., Chen, X., Cheng, K., Li, Y., and Sun, B. (2021). Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 387–396. Springer.

Ma, Z., Luo, M., Guo, H., Zeng, Z., Hao, Y., and Zhao, X. (2024). Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821.

Madan, N., Møgelmose, A., Modi, R., Rawat, Y. S., and Moeslund, T. B. (2024). Foundation models for video understanding: A survey. *Authorea Preprints*.

Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., and Bartoli, A. (2016). Computer-aided classification of gastrointestinal lesions in regular colonoscopy. *IEEE transactions on medical imaging*, 35(9):2051–2063.

Misawa, M., Kudo, S.-e., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al. (2021). Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal endoscopy*, 93(4):960–967.

Nasirihaghighi, S., Ghamsarian, N., Husslein, H., and Schoeffmann, K. (2024). Event recognition in laparoscopic gynecology videos with hybrid transformers. In *MultiMedia Modeling (MMM 2024)*, pages 82–95. Springer.

Nwoye, C. I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., and Padoy, N. (2022). Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433.

Prebay, Z. J., Peabody, J. O., Miller, D. C., and Ghani, K. R. (2019). Video review for measuring and improving skill in urological surgery. *Nature Reviews Urology*, 16(4):261–267.

Schmidgall, S., Kim, J. W., Jopling, J., and Krieger, A. (2024). General surgery vision transformer: A video pre-trained foundation model for general surgery. *arXiv preprint arXiv:2403.05949*.

Smedsrud, P. H., Thambawita, V., Hicks, S. A., Gjestang, H., Nedrejord, O. O., Næss, E., Borgli, H., Jha, D., Berstad, T. J. D., Eskeland, S. L., Lux, M., Espeland, H., Petlund, A., Nguyen, D. T. D., Garcia-Ceja, E., Johansen, D., Schmidt, P. T., Toth, E., Hammer, H. L., de Lange, T., Riegler, M. A., and Halvorsen, P. (2021). Kvasir-Capsule, a video capsule endoscopy dataset. *Scientific Data*, 8(1):142.

Tong, Z., Song, Y., Wang, J., and Wang, L. (2022). Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., and Padoy, N. (2016). Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97.

Wagner, M., Müller-Stich, B.-P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D. M., Müller, B., Davitashvili, T., Capek, M., Reinke, A., Yu, T., Vardazaryan, A., Nwoye, C. I., Padoy, N., Liu, X., Lee, E.-J., Disch, C., Meine, H., Xia, T., Jia, F., Kondo, S., Reiter, W., Jin, Y., Long, Y., Jiang, M., Dou, Q., Heng, P. A., Twick, I., Kirtac, K., Hosgor, E., Bolmgren, J. L., Stenzel, M., von Siemens, B., Kenngott, H. G., Nickel, F., von Frankenberg, M., Mathis-Ullrich, F., Maier-Hein, L., Speidel, S., and Bodenstedt, S. (2021). Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark.

Wang, G., Bai, L., Wang, J., Yuan, K., Li, Z., Jiang, T., He, X., Wu, J., Chen, Z., Lei, Z., et al. (2025). Endochat: Grounded multimodal large language model for endoscopic surgery. *arXiv preprint arXiv:2501.11347*.

Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., and Qiao, Y. (2023a). Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560.

Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al. (2024). Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer.

Wang, Z., Liu, C., Zhang, S., and Dou, Q. (2023b). Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer.

Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.-H., Dou, Q., and Liu, Y. (2022). Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 486–496. Springer.

Yu, T., Mutter, D., Marescaux, J., and Padoy, N. (2018). Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition. *arXiv preprint arXiv:1812.00033*.

Zeng, Z., Luo, M., Kong, X., Liu, H., Guo, H., Yang, H., Ma, Z., and Zhao, X. (2024). Mitigating world biases: A multimodal multi-view debiasing framework for fake news video detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6492–6500.

Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al. (2025). Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Zhao, L., Gundavarapu, N. B., Yuan, L., Zhou, H., Yan, S., Sun, J. J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al. (2024). Videoprism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217*.

Zhou, C., Jiang, R., Luan, F., Meng, S., Wang, Z., Dong, Y., Zhou, Y., and He, B. (2025a). Dual-arm robotic fabric manipulation with quasi-static and dynamic primitives for rapid garment flattening. *IEEE/ASME Transactions on Mechatronics*.

Zhou, C., Xu, H., Hu, J., Luan, F., Wang, Z., Dong, Y., Zhou, Y., and He, B. (2025b). Ssfold: Learning to fold arbitrary crumpled cloth using graph dynamics from human demonstration. *IEEE Transactions on Automation Science and Engineering*.

A 数据集许可证

SurgBench 汇集了多个公开可用的手术视频数据集，并仔细审查了各自的许可，以确保研究界的合规使用。SurgBench 的核心组件，旨在进行直接基准测试和微调，主要由那些允许学术和非商业研究的许可证释放的数据集组成，例如 Cholec80（CC-BY-NC-SA 4.0），Kvasir-Capsule（自定义学术/教育用途），和 SUN-SEG（自定义非商业研究用途基于 Apache 许可证）。值得注意的是，与 SurgBench 一同开发的基础模型在预训练时使用了更广泛的数据集，包括一些限制更多的条款的数据集，如 AVOS，SimSurgSkill2021 和 AIxSuture（CC-BY-NC-ND 4.0），由于许可限制（例如，特定挑战，无衍生作品），这些特定数据集并不作为 SurgBench 微调套件的一部分重新分发。我们强烈建议用户查阅原始资源，以获取每个构成数据集的完整许可详细信息。

B 其他三个类别的饼图

摄像机运动、器官检测、工具识别的饼状图如图 6 所示。

预训练数据的组成如表 6 所示，

72 个标签到任务的映射如表 7 和 8 所示。

C 每个数据集处理步骤的详细信息

私密数据本研究使用的数据集来源于腹腔镜右半结肠切除术的操作，主要关注于手术阶段的分类。原始数据集由来自中国三家三级甲等医院的外科视频构成。这些视频最初以其原始格式存储，并根据腹腔镜右半结肠切除术的能力评估工具（CAT）所概述的不同手术阶段构建的层次目录结构进行组织。主要数据组件包括六个不同的阶段：1. 腹部手术环境的建

Dataset Name	License Type	Usage Conditions
Datasets for SurgBench-P and SurgBench-E		
Cholec80	CC-BY-NC-SA 4.0	Attribution required, non-commercial use, share-alike
Hyper-Kvasir	Custom	Citation required, academic use
Kvasir-Capsule	Custom	Attribution required, research and educational purposes
Colonoscopic Dataset	Custom	Publicly available, registration required
AutoLaparo	CC-BY-NC-SA 4.0	Attribution required, non-commercial use, share-alike
SUN-SEG	Custom	Application required, research and educational purposes
CholecT50	CC-BY-NC-SA 4.0	Attribution required, non-commercial use, share-alike
JIGSAWS	Custom	Application required, academic research, specific citations
Endoscopic Vision 2019	Custom	Form submission required, academic use
Datasets For SurgBench-P Only		
SimSurgSkill2021	Challenge-specific	Limited to challenge scope only
AVOS	Custom protocol	Redistribution prohibited, academic research only
AIxSuture	CC-BY-NC-ND 4.0	Modifications and commercial use prohibited

Table 5: 按许可证对外科数据集进行分类

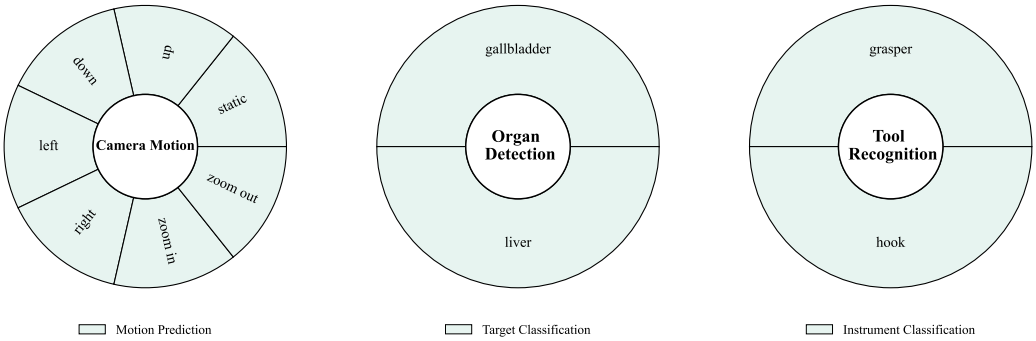


Figure 6: SurgBench-E 中的任务饼图。我们有六个类别，这里显示了三个类别的分布。不同的颜色代表不同的子类别。

Clinical Specialty	Surgical Procedure	Source ID
General Surgery	Appendectomy	S1, S6
General Surgery	Cholecystectomy	S1, S3, S4, S9, S11
General Surgery	Gastrectomy	S1, S6
General Surgery	Hernia Repair	S1, S6
General Surgery	Splenectomy	S1, S7
General Surgery	Pilonidal cystectomy	S1
General Surgery	Bariatric surgery	S1, S6
General Surgery	Anti-reflux surgery	S1, S12
General Surgery	Hepatectomy	S1, S6, S7
Colon and Rectal Surgery	Colectomy	S1, S6, S12, S16
Colon and Rectal Surgery	Colostomy	S1, S6, S12
Thoracic Surgery	Esophagectomy	S1, S6, S12
Urological Surgery	Nephrectomy	S1
Urological Surgery	Adrenalectomy	S1
Obstetrics and Gynecology	Hysterectomy	S1, S10
Obstetrics and Gynecology	Gynecologic Oncology	S1, S10
Neurological Surgery	Neurological Surgery	S1
Ophthalmic Surgery	Ophthalmic Surgery	S1
Oral and Maxillofacial Surgery	Oral and Maxillofacial Surgery	S1
Orthopaedic Surgery	Orthopaedic Surgery	S1
Otolaryngology	Otolaryngology	S1
Pediatric Surgery	Pediatric Surgery	S1

Table 6: 临床专科和外科手术及其来源（ID）的 SurBennch-P

Label	Description
0	surgical_phase-preparation
1	surgical_phase-dividing_ligament_and_peritoneum
2	surgical_phase-dividing_uterine_vessels_and_ligament
3	surgical_phase-transecting_the_vagina
4	surgical_phase-specimen_removal
5	surgical_phase-suturing
6	surgical_phase-washing
7	motion_prediction-static
8	motion_prediction-up
9	motion_prediction-down
10	motion_prediction-left
11	motion_prediction-right
12	motion_prediction-zoom-in
13	motion_prediction-zoom-out
14	instrument_classification-grasper
15	instrument_classification-hook
16	verb_classification-retract
17	verb_classification-dissect
18	target_classification-gallbladder
19	target_classification-liver
20	phase_classification-preparation
21	phase_classification-carlot-triangle-dissection
22	phase_classification-gallbladder-dissection
23	phase_classification-cleaning-and-coagulation
24	phase_classification-gallbladder-extraction
25	disease_classification-adenoma
26	disease_classification-hyperplasic
27	disease_classification-serrated
28	phase_classification-preparation
29	phase_classification-calot_triangle_dissection
30	phase_classification-clipping_and_cutting
31	phase_classification-galbladder_dissection
32	phase_classification-galbladder_packaging
33	phase_classification-cleaning_and_coagulation
34	phase_classification-galbladder_retraction
35	disease_classification-home
36	disease_classification-home
37	disease_classification-home

Table 7: 标签描述（第 1 部分）

立，2. 右结肠和右半结肠后腹膜间隙的解剖，3. 肠系膜上血管的识别和结扎，4. 横结肠和亨利干后腹膜间隙的解剖，5. 胃结肠韧带的开放和肠系膜间隙（IMS）的识别，6. 标本取出和胃肠道重建（腹腔内和腹腔外吻合）。每个视频都被注释了关于曝光、不良事件、技术操作、和质量评估的详细元数据，能够对外科手术的熟练程度进行全面评估。

为了准备机器学习任务的数据集，我们使用 Python 脚本实现了一个系统化的预处理流程。首先对原始视频按 30 秒片段进行分割，然后通过将每个片段加速到 3 倍速来压缩至 10 秒，同时保留基本的视觉信息。这个过程利用了 `ffmpeg` 库，确保了精确的帧级提取和高效的压缩。分割逻辑涉及使用 `ffprobe` 计算视频的总时长，将其划分为不重叠的区间，并为视频和音频流应用统一的滤波链以保持同步。片段以 H.264 编解码器和 AAC 音频编码格式保存，并使用 `+faststart` 标志优化以便于快速网络流传输。每个片段的元数据，包括其标签、相对路径、压缩后的时长以及相关的任务类型，会记录在一个 JSON 文件中以便后续的模型训练。

处理后，数据集被划分为 N 个视频片段，覆盖六个类别，每个片段的标准化时长约为 10 秒。视频分辨率保持为 1920×1080 像素，帧率为每秒 30 帧（fps）。所有片段的总帧数大约为 $30 \times N$ ，为时间建模提供了丰富的资源。私有数据采用分层结构，每个片段都根据其手术

Label	Description
38	disease_classification-home
39	disease_classification-home
40	disease_classification-home
41	disease_classification-home
42	disease_classification-home
43	disease_classification-home
44	gesture_classification-reaching_for_needle_with_right_hand
45	gesture_classification-positioning_needle
46	gesture_classification-pushing_needle_through_tissue
47	gesture_classification-transferring_needle_from_left_to_right
48	gesture_classification-moving_to_center_with_needle_in_grip
49	gesture_classification-pulling_suture_with_left_hand
50	gesture_classification-orienting_needle
51	gesture_classification-using_right_hand_to_help_tighten_suture
52	gesture_classification-dropping_suture_at_end_and_moving_to_end_points
53	gesture_classification-reaching_for_needle_with_left_hand
54	gesture_classification-making_c_loop_around_right_hand
55	gesture_classification-reaching_for_suture_with_right_hand
56	gesture_classification-pulling_suture_with_both_hands
57	disease_classification-ileocecal_valve
58	disease_classification-blood-fresh
59	disease_classification-foreign-body
60	disease_classification-lymphangiectasia
61	disease_classification-normal-clean-mucosa
62	disease_classification-pylorus
63	disease_classification-reduced-mucosal-view
64	polyp_detection-clips_without_polyps
65	polyp_detection-clips_with_polyps
66	phase_classification-1_establishment_of_abdominal_surgical_environment
67	phase_classification-2_dissection_of_the_posterior_peritoneal_space
68	phase_classification-3__identification_and_ligation_of_the_vessels_on_the_mesentery
69	phase_classification-4_dissection_of_the_posterior_peritoneal_space
70	phase_classification-5_opening_of_the_gastrocolic_ligament_and_identification
71	phase_classification-6_specimen_removal_and_gastrointestinal_reconstruction_(intra-abdominal_and_extra-abdominal_anastomosis)

Table 8: 标签描述（第 2 部分）

阶段进行标记，从而能够进行监督学习的阶段识别。包含每个片段详细注释的最终元数据文件以 JSON 格式保存。

AutoLaparo AutoLaparo 是一个多任务数据集，专门为推进腹腔镜子宫切除术中图像引导的手术自动化而设计。原始数据集包括三个主要任务：手术工作流识别（任务 1）、腹腔镜运动预测（任务 2）以及器械和关键解剖结构分割（任务 3）。任务 1 包含 21 段腹腔镜子宫切除术视频，每段视频以每秒 1 帧的频率标注了七个手术阶段，而任务 2 由 300 个视频剪辑组成，每个剪辑时长 10 秒，标注了七种腹腔镜运动标签。任务 3 提供了 1800 张带有相应掩膜的图像，标注了四种类型的手术器械和一个关键解剖结构。数据集被组织成特定的目录：任务 1 的视频和标签分别存储在 `task1/videos` 和 `task1/labels` 中；任务 2 的剪辑及其标签位于 `task2/clips` 和 `task2/laparoscope_motion_label.txt` 中；任务 3 的图像和掩膜存储在 `task3/images` 和 `task3/masks` 中。对于预处理，任务 1 的视频被分割成每段 100 帧的剪辑，以确保一致的剪辑长度，而任务 2 的剪辑直接按提供的使用。数据集被分割为训练、验证和测试集：对于任务 1，视频 01-10 用于训练，11-14 用于验证，15-21 用于测试；对于任务 2 和任务 3，剪辑 001-170 被指定用于训练，171-227 用于验证，228-300 用于测试。后处理完成后，数据集包含 1,388 分钟的视频，帧率为 25 帧每秒，分辨率为 1920 × 1080 像素。这种细致的划分确保了一个平衡且具有代表性的数据集，以便进行稳健的模型训练和评估，促进手术自动化任务的全面基准测试。

Cholec80 Cholec80 数据集源于斯特拉斯堡大学医院/IRCAD，是一个用于外科工作流程分析的腹腔镜胆囊切除手术视频的全面集合。该数据集包含 80 个高分辨率视频，每个视频都有详细的阶段标签和工具存在信息注释，是外科视频理解研究的宝贵资源。数据集的主页（`http://camma.u-strasbg.fr/datasets`）提供了其应用的概述，包括阶段识别和工具检测任务。此外，随附的 README 文件和注释提供了对数据结构、许可和引用要求的见解。

最初，数据集包含 80 个以每秒 25 帧（fps）录制的视频，视频的分辨率各不相同。每个视频都附有两种类型的注释：阶段注释和工具注释。阶段注释为七个手术阶段提供帧级标签，包括准备阶段、Calot 三角解剖、夹钳和切割、胆囊解剖、胆囊包装、清洁和凝固、以及胆囊牵引。工具注释，每秒采样 1 帧，指示七种手术工具的存在或不存在，比如夹持器、双极设备、钩子、剪刀、剪切器、冲洗器和标本袋。这些注释存储在制表符分隔的文本文件中，包括帧索引和相应的标签。此外，提供了带有时间戳的阶段注释，以便在视频回放过程中进行可视化。

为了实现高效的处理和分析，数据集被预处理成较短的视频片段。每个视频被分成大约 300 帧的不重叠片段，在原始帧率为 25 fps 的情况下，每个片段对应于 12 秒。这种分割确保了片段的持续时间易于管理，同时为下游任务保留了足够的时间上下文。对于每个片段，分析了阶段和工具标注以确定主要标签。具体来说，一个片段中最频繁出现的阶段标签被作为片段的阶段分类标签，而工具的出现则是根据各帧中每个工具的最大出现次数来确定的。生成的片段以 MP4 格式保存，使用 FFmpeg，保持了原始分辨率和帧率。预处理管道还生成了 JSON 格式的元数据，详细说明了片段路径、持续时间和相关标签。

经过预处理，数据集由大约 670 个视频片段组成，总时长约为 80 分钟。每个片段保留了原始的分辨率和 25 fps 的帧率，以确保与原始视频的一致性。平均片段长度为 12 秒，由于片段边界与视频长度的对齐，存在轻微的变化。所得数据集非常适合用于训练和评估外科手术阶段识别和工具存在检测模型，提供了外科手术阶段和工具使用模式的平衡分布。这种结构化表示有助于开发针对腹腔镜视频分析的深度学习模型，推动计算机辅助外科手术的发展。

Colonoscopic-web Colonoscopic-Web 数据集包含来自结肠镜检查的 76 个视频，记录了不同类型的胃肠道病变。每个视频都伴有注释的真实数据，包括组织病理学发现、专家注释和记录系统的校准数据。数据集中包含腺瘤、增生性和锯齿状三个类别，代表常见息肉类型的分布。较长的视频被剪成 2 分钟的短视频。该数据集旨在用于自动病变检测和分类的研究，目的是通过减少常规检查中对染色内镜的依赖，提高诊断准确性，减少临床医生的工作负担。

Endovis2019 Endovis2019 数据集包含来自普通外科手术室的内窥镜视频数据，这些数据是在海德堡大学医院及其附属医院进行腹腔镜手术时获得的。所有手术都由外科专家逐帧标记，标签包括手术阶段、外科动作和使用的器械。记录的手术主要是腹腔镜胆囊切除术。该数据集包含来自至少三家医院的 30 台不同手术的录制视频，并提供了每个手术过程的内窥镜视频。为确保隐私，所有额外拍摄到的腹部以外的镜头都被完全白色的帧遮盖。该数据集的任务包括阶段识别、动作分类和工具分类。阶段识别任务涉及基于手术过程的不同阶段进行分类。数据集根据从准备阶段到胆囊牵引的多个不同阶段进行标注。动作分类任务涉及对外科操作动作进行分类，包括抓取、持握、切割和夹持动作。工具分类任务需要对手术中使用的各种工具进行分类，涵盖了广泛的手术器械，如抓持器、夹子、凝血器械、剪刀、吸引-冲洗器、样本袋和订书机。此外，还定义了一个“未定义器械杆”类别，用于标记未明确分类的工具。基于每一帧的标签，视频通过帧分类进行剪切，具有相同标签的连续帧形成一个新的视频段，然后每个新的视频段被剪切为 10 秒的视频。

hyper-kvasir Hyper-Kvasir 数据集是一个包含来自挪威卑尔根医院的 374 个胃肠镜和结肠镜视频的胃肠道图像和视频数据集，总计 11.62 小时，超过 100 万帧。该数据集根据分级结构进行了分类。首先，该数据集在第一层级上用上消化道（Upper GI）和下消化道（Lower GI）进行分类。然后根据四个主要类别进行进一步细分，包括解剖标志、病理发现、粘膜视野质量以及治疗性干预，形成第二层级。这些类别进一步细分为 30 个属于第三层级的具体子类别。通过将这些子类别与上消化道和下消化道的标签结合，每个视频被赋予一个新标签。每个长视频根据其分类被剪切成 5 秒的小视频片段，并存储在相应的文件夹中。此外，该数据集包含 10,662 张标注图像，这些图像以 JPEG 格式存储，并以 23 种不同的病变类型进行标注，覆盖了胃肠道不同部分的广泛正常和病理表现。这些图像和视频为开发 AI 辅助胃肠内镜分析系统提供了大量训练数据，特别是可以帮助研究人员应对医学数据中常见的类别不平衡问题。

kvasir-capsule Kvasir-Capsule 数据集是一个用于胃肠内窥镜视频分析的开放数据集。该数据集包含 4,741,621 条数据，包括 47,238 张带有边界框的标记图像、43 个标记视频和 74 个未标记视频。此外，可以从所有视频中提取出 4,694,266 张未标记图像。数据集中的标记图像

是从原始长视频中的帧中提取出来的。基于这些标记帧的信息，原始视频被切割成每 5 秒一个的短视频，并与原始长视频的标签保持一致。数据集中有 47,238 张标记图像，这些图像被分类为 14 个不同类别，并分为两个主要类别：解剖结构和腔内发现。解剖结构包括与胃肠道相关的解剖学标记，例如幽门、回盲瓣以及 Vater 壶腹，而腔内发现则涵盖胃肠道发生的病理变化，例如正常干净黏膜、黏膜视野减退、新鲜血、血红蛋白等。这些图像广泛用于研究胃肠疾病，并帮助提高内窥镜诊断的准确性。同时，由于数据集中各种疾病图像数量的不平衡，特别是对于一些较为罕见的疾病，研究人员需要采用有效的机器学习方法，以便能够从有限的训练数据中学习，尤其是对罕见疾病的识别。

CholecT50 CholecT50 是一个用于细粒度动作识别的腹腔镜胆囊切除术内窥镜视频数据集，旨在促进腹腔镜手术中动作识别技术的研究。该数据集包含来自 Cholec80 数据集的 45 个视频和内部数据集 Cholec120 的 5 个视频，所有视频均具有每个手术动作的三重信息（< 器械, 动词, 目标 >）的精细标签及相应的阶段标签。该数据集还提供了 5 个视频中的器械尖端的空间注释（边界框）以及所有视频中的帧-三重匹配标签。此外，每个视频中的帧以每秒 1 帧的频率提取，并注释了详细的手术动作，旨在支持腹腔镜手术中运动识别算法的研究。数据集被分成四个任务：阶段分类、器械分类、动词分类和目标分类。任务划分基于每帧的 JSON 注释文档，每帧都详细标注，几帧连续帧具有相同的标签。为了适应模型训练，所有视频片段被分割成数个 5 秒的短视频，帧率为 10，并重新排列标签，将四个任务的标签映射到 0-31 的范围内，以便统一标准化训练数据。

LDPolyVideo LDPolypVideo 数据集是一个用于结肠镜息肉检测的大型多样化数据集。它包含从常规的临床结肠镜检查中收集的结肠镜视频，并去除了所有与患者相关的元数据。该数据集包含 160 个视频，总计 40,266 帧，其中 33,884 帧包含息肉。为了增加数据集的多样性，数据不仅包括清晰的息肉图像，还涵盖了由于摄像机运动造成的运动模糊，肠道折叠和肠道蠕动。数据集被分为训练、验证和测试集。验证集和测试集中的图像被标记以指示每帧是否包含息肉，分别用标记值 0 或 1 表示。对于这些帧，视频被合成为 fps 为 25 的视频剪辑，并根据帧级别标签剪切成 5 至 10 秒的短片。然后，训练集根据整个视频的标记分为包含息肉的视频和不包含息肉的视频。对于包含息肉的视频，视频平均分为 2 至 4 段，每段被压缩为 10 秒钟的视频，以确保每段中都包含息肉。对于不包含息肉的视频，视频被剪切成 5 到 10 秒的短片。这种处理方法确保每个片段被正确标记，提升了数据集的多样性和模型的泛化能力。

JIGSAWS JIGSAWS 数据集用于建模人体运动中的手术活动，由约翰霍普金斯大学与直观外科公司合作收集。该数据集使用达芬奇手术系统进行收集，涉及八名不同技能水平的外科医生，他们在一个桌面模型上执行三个基本手术任务的五次重复：缝合、打结和穿线，这些是大多数手术技能训练课程的标准组成部分。JIGSAWS 数据集由三个主要组成部分构成：第一个组成部分是运动学数据，包括描述操作者运动的笛卡尔位置、方向、速度、角速度和描述操纵器运动的夹角；第二个组成部分是使用立体视频的视频数据；第三个组成部分是手动标记的数据，包括手势和技能。数据集总共包含 15 个标记手势，每个手势对应于手术过程中的特定任务段。具体的 15 个标签包括 G1：" 用右手抓住针"、G2：" 定位针"、G3：" 将针穿过组织"、G4：" 将针从左手传递到右手"、G5：" 抓住针并移动到中心"、G6：" 用左手拉紧缝线"、G7：" 用右手拉紧缝线"、G8：" 调整针的方向"、G9：" 用右手帮助收紧针脚"、G10：" 放松更多针脚"、G11：" 丢弃针脚并移动到最后"、G12：" 用左手抓住针"、G13：" 在右手周围做一个 C 形"、G14：" 用右手抓住缝边"、G15：" 用双手拉动针脚"。这些标记信息存储在每个任务的转录文件中，包括穿线、缝纫和系结。

AIxSuture AIxSuture 数据集是为了分析通过虚拟现实头戴显示器指导训练的效果而收集的。它包含 314 个视频，记录了学生在模拟环境中进行外科缝合的过程，并根据技能评分将他们分为熟练者、新手和中级者类别。技能通过技术技能客观结构化评估量表进行评分，共有八个技能类别，所有类别评分的总和形成一个范围从 8 到 40 的总体评分 GRS。通过对三位评估者评分之间的成对皮尔逊相关系数的初步分析获得平均值。对于每段视频，三位评估者的评分取平均值，然后将其分为新手、中级和熟练者三个类别。由于视频时间较长，需要对长视频进行理解，这些视频被均分为 2 到 4 个部分，并最终压缩成 10 秒段进行处理和分析。

AVOS 标注的开放式手术视频（AVOS）收集了从 YouTube 上获取的 1997 个开放手术视频，涵盖了过去 15 年内 23 种手术类型和 50 个国家。其中，有 326 个 [3] 视频每五秒对场景动作进行了标注。我们首先从数据集的主页（`https://research.bidmc.org/surgical-informatics/avos`）下载数据，将标注视频分割成每 5 秒的短片段，保留原始帧率和分辨率，并有五个动作标签（切割、缝合、打字、无动作和背景）。最终获取了 27,745

[3] 30 个视频网址已过期，所以我们只有 296 个标注视频

个片段，总计 38.53 小时和 3,582,276 帧。由于视频来源多样，其帧率变化不一，主要的 FPS 分布集中在 25 到 30 之间。