SurgPub-Video: A Comprehensive Surgical Video Framework for Enhanced Surgical Intelligence in Vision-Language Model

Yaoqian Li[1]*, Xikai Yang[1]*, Dunyuan Xu[1]*, Yang YU[1], Litao Zhao[1], Xiaowei Hu[2], Jinpeng Li[1]†, Pheng-Ann Heng[1,3]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
[2]School of Future Technology, South China University of Technology, Guangzhou, China
[3]Institute of Medical Intelligence and XR, The Chinese University of Hong Kong, Hong Kong, China
{yqli, xkyang22, dyxu21, jpli21}@cse.cuhk.edu.hk

Abstract

Vision-Language Models (VLMs) have shown significant potential in surgical scene analysis, yet existing models are limited by frame-level datasets and lack high-quality video data with procedural surgical knowledge. To address these challenges, we make the following contributions: (i) SurgPub-Video, a comprehensive dataset of over 3,000 surgical videos and 25 million annotated frames across 11 specialities, sourced from peer-reviewed clinical journals, (ii) SurgLLaVA-Video, a specialized VLM for surgical video understanding, built upon the TinyLLaVA-Video architecture that supports both video-level and frame-level inputs, and (iii) a video-level surgical Visual Question Answering (VQA) benchmark, covering diverse 11 surgical specialities, such as vascular, cardiology, and thoracic. Extensive experiments, conducted on the proposed benchmark and three additional surgical downstream tasks (action recognition, skill assessment, and triplet recognition), show that SurgLLaVA-Video significantly outperforms both general-purpose and surgical-specific VLMs with only three billion parameters.

Datasets — https://github.com/Yaoqian-Li/SurgPub-Video

Introduction

Vision-Language Models (VLMs) have demonstrated exceptional performance across various domains (Liu et al. 2023; Li et al. 2023). Recent work in surgical knowledge integration has shown promising results on tasks such as scene understanding, action recognition, and clinician skill assessment (Zeng et al. 2025; Liu et al. 2025). However, current surgical VLMs are fundamentally constrained by their exclusive training on frame-level datasets (Seenivasan et al. 2022; Yuan et al. 2024a), thereby lacking the ability to understand temporal dynamics in surgical workflows.

This limitation primarily stems from the scarcity of high-quality surgical video datasets with instruction-following annotations, hindering the advancement of surgical VLMs. The curation of such datasets faces three key obstacles: (i) **Limited video diversity**: While frame-level surgical VQA datasets claim a large number of frames, they derive from a small pool of source videos, failing to capture the representative spectrum of surgical scenarios. (ii) **Insufficient annotation granularity**: Surgical videos typically contain only sparse categorical labels, and converting them into detailed instruction-following samples requires extensive man-

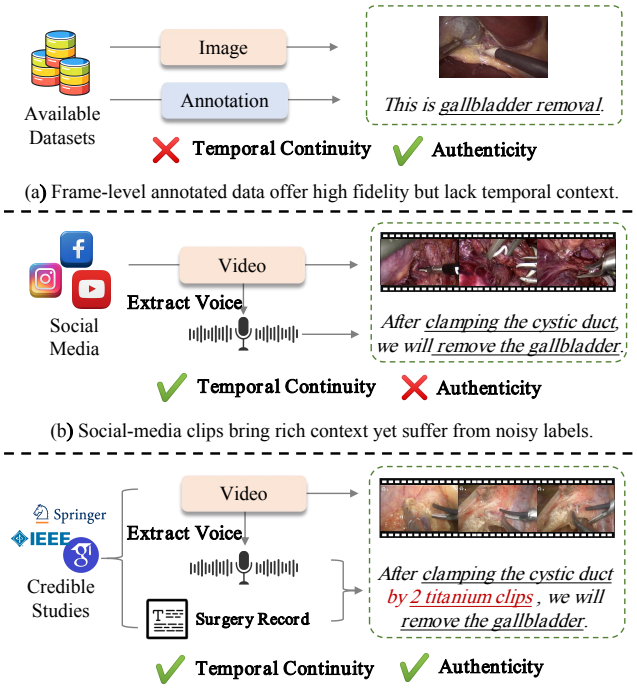


Figure 1: Comparison of three pipelines for building surgical-video VQA datasets. (a) Converting public frame-label datasets offers high annotation fidelity but lacks narrative context. (b) Mining social media, e.g., YouTube, video narration supplies rich temporal descriptions yet suffers from inconsistent accuracy without peer review. (c) Leveraging peer-reviewed journal videos combines continuous operative narration with authoritative reports, providing both temporal coherence and professional reliability.

*These authors contributed equally.
†Corresponding author

Dataset	Year	QA Level	Richness		Scale		Multi-Center
			#Surgery Types	Language Source	#Videos	#Frames	
Cholec80-VQA	2022	Image	1	Categorical Text	40	21.6K	No
EndoVis-18-VQA	2022	Image	1	Categorical Text	19	2K	No
EndoVis-VQLA	2023	Image	1	Categorical Text	29	2.2K	No
PSI-AVA-VQA	2024	Image	1	Categorical Text	8	2.2K	No
SSG-VQA	2024	Image	1	Categorical Text	50	25.5K	No
Surg-396K	2024	Image	3	Categorical Text	109	41.4K	Yes
SurgVLM-DB	2025	Image	16	Categorical Text	849	1.81M	Yes
SurgPub-Video (Ours)	2025	Video	75	Transcript & Report	3,538	25M	Yes

Table 1: Comparison of our SurgPub-Video with other surgical VQA datasets. Our SurgPub-Video is the only dataset specifically designed to support video-level surgical VQA, offering a unique advantage of temporal surgical scene understanding.

ual effort and is error-prone. (iii) **Unreliable data sources**: Videos sourced from social media often lack credibility and accuracy, rendering them clinically unreliable.

Recent studies have explored a range of strategies to mitigate these challenges. To address the issue of data scarcity, some works attempt to collect existing public datasets and use frame-level annotations to construct surgical VQA datasets (Zeng et al. 2025; Liu et al. 2025). However, compared with native video-level labeling, converted annotations lack crucial temporal information, restricting the procedural understanding ability of surgical VLMs. Furthermore, only a limited number of surgical databases are entirely composed of video-based instances, and these often focus on pretraining image encoders using video captions (Yuan et al. 2025, 2024b). Consequently, these datasets do not include comprehensive question-answer pairs, which are essential for fine-tuning VLMs and improving their understanding of surgical tasks. Another key challenge in current datasets is the lack of rich textual information for generating versatile QA pairs. Since current image-level surgical datasets (Seenivasan et al. 2022; Bai et al. 2023; Seenivasan et al. 2023) usually provide only coarse category labels, such as "gallbladder removal" or "liver" for QA generation, omitting semantically rich descriptions of the visual scene. This limitation leads to VLMs aligning visual data with discrete labels, rather than with the open, descriptive space of natural language. For video-type data, a practical alternative is to leverage automatic speech recognition (ASR) models to generate transcripts as textual information (Yuan et al. 2025). Nevertheless, ASR outputs often contain noise and inconsistencies, requiring extensive post-processing. Furthermore, the reliability of ASR transcripts is questionable, especially for videos not reviewed by clinicians, such as those sourced from social media platforms (e.g. YouTube) (Yuan et al. 2024c; Che et al. 2025), where accuracy and authenticity are uncertain without expert validation.

To address these challenges, in this study, we construct the SurgPub-Video dataset, a high-quality surgical VQA dataset composed entirely of videos. As shown in Fig.1, unlike previous surgical VQA datasets, our SurgPub-Video dataset exhibits superior qualities in terms of both temporal continuity and authenticity. Specifically, we restrict our crawling scope to original surgical videos published in peer-reviewed articles, ensuring inherent vetting by the academic and clinical community for trustworthiness. As shown in Tab.1, we collect a total of 3,538 surgical videos, significantly surpassing the size of previous surgical VQA datasets. To create a comprehensive and high-quality surgical VQA database entirely composed of video clips, we move beyond categorical labels by utilizing both audio transcripts and associated surgical records to extract precise and semantically rich information. We also design a sophisticated dataset curation workflow, including audio-guided video clip preprocessing, structured surgical concept extraction, and human-involved VQA pair creation. The final SurgPub-Video dataset contains 10,926 surgical clips and 48,520 VQA pairs, spanning 1,823 anatomical structures. Moreover, considering current surgical VLMs heavily rely on the naive LLaVA structure, which only supports frame-level input and lacks explicit temporal modeling, to bridge this gap, we further introduce the SurgLLaVA-Video model, built upon the TinyLLaVA-Video architecture (Zhang et al. 2025), which naturally supports whole-video input and leverages a resampler to integrate in-context relationships across the video. Finally, to comprehensively evaluate the capabilities of VLMs in surgical video understanding, we also propose the SurgPub-Video benchmark derived from our dataset, covering 11 vital surgical specialties and 5 main surgical tasks. Our contribution can be summarized as follows:

- We introduce a high-quality, large-scale surgical dataset, SurgPub-Video, for VLM training, which contains VQA pairs with good temporal continuity, semantic richness, and authenticity.
- We develop the SurgLLaVA-Video, which is trained on SurgPub-Video and enables both video-level and frame-level surgical scene understanding.
- We construct an extensive benchmark encompassing various surgical procedures and tasks. We systematically evaluate both open-source and commercial multimodal models on this benchmark and find that our SurgLLaVA-Video model achieves superior performance compared to existing dedicated and closed-source models.

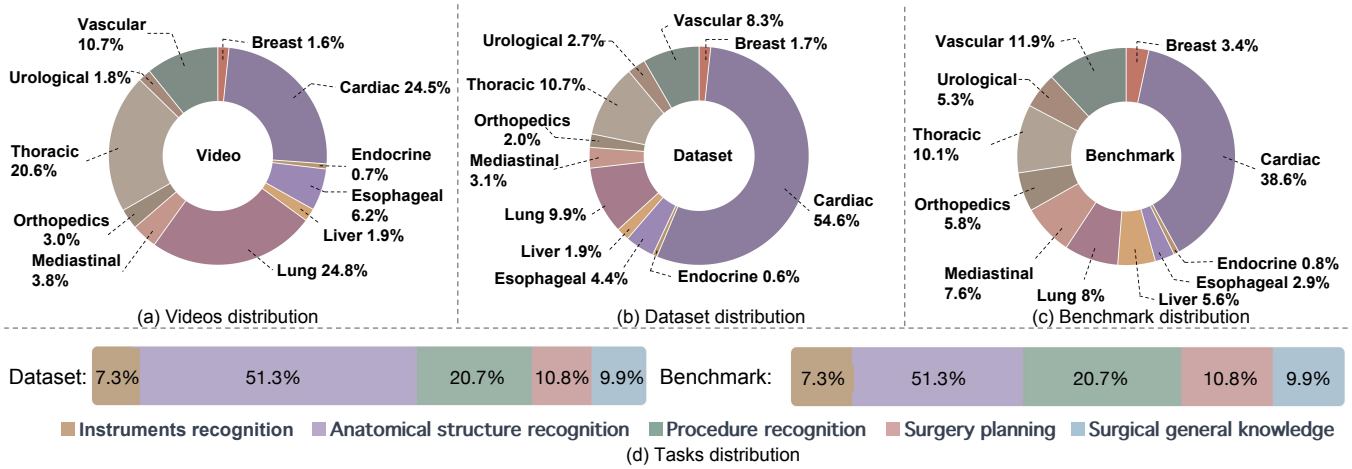


Figure 2: Overview of the SurgPub-Video data distribution. (a) The proportion of original journal videos in 11 surgical specialties; (b) The specialty distribution of the SurgPub-Video; (c) The distribution of Benchmark specialties; (d) The coverage ratio of five types of VQA tasks (instrument, anatomy, procedure, planning, and general knowledge) in the Benchmark and complete datasets, demonstrating the credibility of data sources, the balance of specialties, and the diversity of tasks.

Related Work

Surgical Multimodal Datasets and Benchmarks

A large-scale surgical VQA dataset is crucial for enabling vision–language models (VLMs) to comprehend complex surgical scenes. Yet, traditional single-task image-level datasets with limited annotations cannot support diverse organs or fine-grained reasoning. Recent works address this by converting small datasets, merging multiple sources, or mining web videos. Cholec80-VQA (Seenivasan et al. 2022) and SSG-VQA (Yuan et al. 2024a) reformulate single-task datasets into VQA pairs but lack annotation diversity. SurgVLM-DB (Zeng et al. 2025) and EndoBench (Liu et al. 2025) expand coverage by aggregating datasets across procedures and organs, while Surg-3M (Che et al. 2025), SurgVLP (Yuan et al. 2024c), and SurgVISTA (Yang et al. 2025) leverage web-sourced videos to scale VQA construction. However, manually curated datasets (Zeng et al. 2025; Liu et al. 2025) still cover limited scenarios, and web-collected data often lack peer-reviewed quality, restricting VLM performance. We tackle these issues by building a comprehensive VQA dataset from high-quality academic surgical videos.

Surgical Large Vision–Language Models

General VLMs, such as LLaVA (Liu et al. 2023), Qwen-VL (Bai et al. 2025), and CogVLM (Wang et al. 2024), achieve strong multimodal reasoning through large-scale instruction tuning. Medical VLMs, including Med-PaLM (Tu et al. 2024), OmniMedVQA (Hu et al. 2024), and LLaVA-Med (Li et al. 2023), specialize on domain data such as MIMIC-CXR (Johnson et al. 2019) and SLAKE (Liu et al. 2021) for diagnostic and interpretive VQA. Surgical VLMs, including SurgVLM (Zeng et al. 2025), SurgicalGPT (Seenivasan et al. 2022), and EndoChat (Wang et al. 2025), adapt these methods for single-frame tasks like phase or tool recognition. Yet, without large, scenario-rich surgical video VQA datasets, they remain limited to static imagery. We bridge this gap with SurgLLaVA-Video, trained on the SurgPub-Video dataset, introducing video-level reasoning into surgical VLMs.

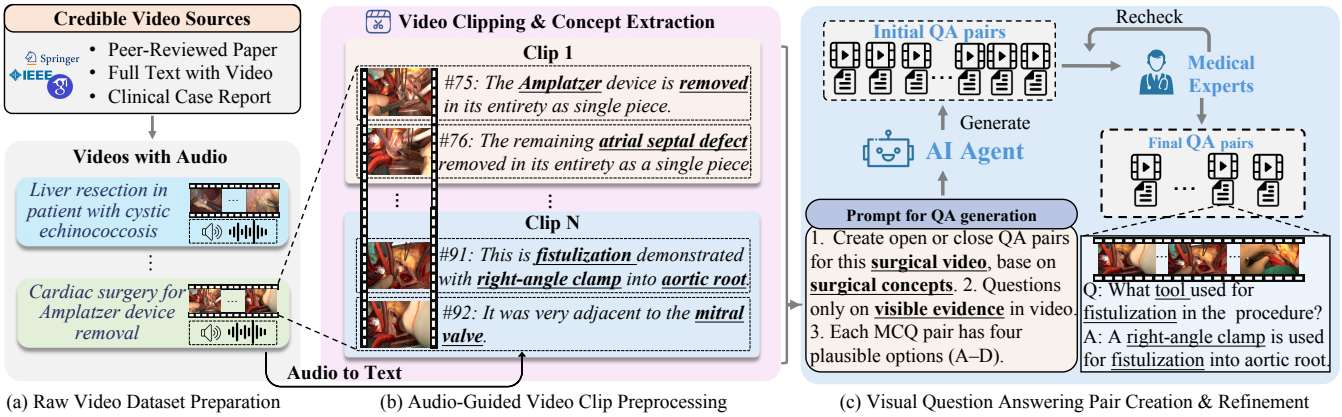
Methodology

In this section, we first present our constructed SurgPub-Video dataset and benchmark, followed by an introduction to the dataset curation workflow. Finally, we provide a description of our SurgLLaVA-Video model architecture.

SurgPub-Video Dataset & Benchmark

SurgPub-Video Dataset Tab. 1 compares our constructed SurgPub-Video database against several popular surgical VQA databases. Unlike most surgical databases that only provide single-frame snapshots, SurgPub-Video targets video-level analysis by consisting entirely of video clips, thereby preserving crucial temporal relationships that are essential for understanding surgical procedures. Fig. 2 shows the distributions of the original video source, SurgPub-Video dataset, and the benchmark. Specifically, SurgPub-Video consists of 10,926 surgical video clips with 29.85s average duration and 48,520 open and closed VQA pairs generated from five concept categories (instruments, procedural, anatomical structure, planning, general surgical knowledge). The clips span 11 surgical specialties and cover 75 kinds of surgery, 1,823 anatomical structures, 40 surgical instruments, and 1,290 unique procedures.

SurgPub-Video Benchmark We randomly sample 20% of VQA pairs from 705 videos to construct the benchmark. To mitigate data imbalance and fairly evaluate the model's performance on different tasks, we remove some similar VQA pairs from the majority category. The final benchmark subset consists of 3,337 samples, reducing cardiac surgery



(a) Raw Video Dataset Preparation (b) Audio-Guided Video Clip Preprocessing (c) Visual Question Answering Pair Creation & Refinement

Figure 3: VQA-generation pipeline for SurgLLaVA-Video dataset and benchmark. (a) Peer-reviewed journal videos and accompanying reports are collected. (b) Whisper transcripts and semantic filtering split each recording into coherent clips and extract core surgical concepts. (c) A prompt-based LLM agent produces initial open-ended and multiple-choice QA pairs for every clip; medical experts then review and refine these drafts, yielding the final high-quality VQA dataset.

samples to 38.6%, and enhancing vascular and mediastinal surgeries to 11.9% and 7.6%, respectively. The benchmark assesses model performance through five VQA tasks, including instrument recognition, anatomical structure recognition, procedural step identification, surgical planning, and general surgical knowledge. The training set and test set are strictly video-level separated. Evaluation metrics include overall accuracy, specialty-specific accuracy, and task-specific accuracy, enabling detailed analysis of model capabilities and limitations across different surgical specialties and different tasks.

Dataset Curation Workflow

To construct a reliable surgical video-based VQA dataset, we collect surgical videos from more than 25 peer-reviewed medical journals. Inspired by surgeons' VQA construction workflow, we design an LLM-assisted, coarse-to-fine pipeline that (i) transcribes audio, (ii) filters and merges clips, (iii) extracts structured concepts, and (iv) generates and refines VQA pairs. This pipeline progressively removes redundancy, preserves clinically relevant context, and produces clip-level samples with paired textual evidence for VQA creation. The detailed pipeline is described as follows.

Raw Video Dataset Preparation We crawl original videos from 25 peer-reviewed medical journals (listed in the supplement) and retain articles accompanied by surgical videos, yielding 3,538 raw surgical videos. During initial filtering, we select articles with surgical videos and exclude entries without usable video content. Fig. 2(a) summarizes the sources: 11 surgery specialties, 323.5 seconds average duration, 1920*1080 resolution, and 30Hz frame rate. Compared to web videos (e.g., YouTube), each video is paired with a peer-reviewed article, improving reliability and providing sufficient accompanying text. We also preserve the associated article text as auxiliary context for later concept extraction and QA prompting.

Audio-Guided Video Clip Preprocessing Considering the redundancy in original surgical videos, we adopt an audio-guided approach for coarse-level clip selection. We employ an audio-to-text (ATT) agent powered by the OpenAI Whisper (Radford et al. 2023) model to generate timestamped transcripts. The ATT agent segments the full audio stream into short segments and converts them to text; however, the corresponding video fragments can be overly fragmented, and many correspond to non-surgical content. To address this, we deploy a coarse-grained video processing (CGVP) agent to filter and integrate these fragments into video clips ranging from 15 to 30 seconds. CGVP removes redundant/non-surgical segments based on transcript content (e.g., non-surgical portions at the beginning/end of videos) and merges adjacent fragments with similar or contextually related semantics. This step reduces fragmentation while preserving temporal coherence within each clip. This merging is important because surgical procedures often span multiple sequential steps. For example, vessel dissection may include fat clearance, hemostatic clamp placement, and vessel incision; CGVP combines related fragments to better capture complete procedural units. Fig. 3 shows the detailed process of generating clips from raw video.

While CGVP removes most redundancy, some clips may still contain unrelated information or unwanted concepts. We therefore deploy a fine-grained concept extraction (FGCE) agent to determine whether clips contain concepts of interest. We categorize surgical concepts into five types: instruments, procedural steps, anatomical structures, treatment planning, and general surgical knowledge. FGCE prompts the LLM with each clip transcript and the corresponding article text to classify the clip and organize the database in a structured format, recording segment length, text, and included concepts. This structured organization supports downstream QA prompting by providing explicit concept tags and the associated evidence context.

Visual Question Answering Pair Creation & Refinement After clip cleaning, a QA generation (QAG) agent selects video clips and their corresponding textual information to construct QA pairs via targeted prompting. QAG sequentially processes clips and associated context, producing initial QA drafts that emphasize clinically meaningful reasoning. To enhance clinical relevance, we design reasoning-based questions derived from causal explanations and rationales stated in video narration, covering prediction of surgical outcomes, operative rationale, and risk-related decision making. For answer generation, the agent produces both open-ended responses and multiple-choice questions, and designs distractors to reduce guessability. We store each QA pair together with its clip and supporting evidence to facilitate supervised training. All answers are kept medically accurate, concise, and aligned with textual evidence. We provide a prompt abstract and an example QA pair in Fig. 3. We involve human expert review to refine the generated VQA pairs, ensuring medical correctness and semantic consistency with the original intent. This review can be iterative, progressively improving data quality in both the SurgPub-Video dataset and the benchmark.

Architecture of SurgLLaVA-Video Model

Here, we present the architecture of our trained SurgLLaVA-Video model. Distinct from the recently proposed SurgVLM (Zeng et al. 2025), which concentrates on frame-level analysis and directly concatenates visual tokens and text tokens as input to the LLM, SurgLLaVA-Video is particularly optimized for video-level input, which comprises three main components: a vision encoder, a video group resampler, and a LLM, following (Zhang et al. 2025). Fig. 4 presents the model architecture, where the vision encoder first receives a video clip $X \in \mathbb{R}^{T\times H\times W\times 3}$ as input and extracts critical visual features at the frame level $Z \in \mathbb{R}^{T\times N\times D}$. T denotes the video clip length, H and W represent the height and width of each frame, respectively, N is the number of tokens corresponding to a single frame, and D is the dimension of the extracted feature embedding. The extracted features Z are then reshaped into $\hat{Z} \in \mathbb{R}^{(T\times N)\times D}$ for downstream operations. Unlike the regular LLaVA structure, SurgLLaVA-Video incorporates a video resampler $\mathbf{P}_\theta$ that dynamically projects visual information into a fixed number of learnable queries $Q \in \mathbb{R}^{L\times D}$, simultaneously maintaining temporal relations across frames while saving computational resources. Inside the video resampler, due to the limited number of queries, we divide $\hat{Z}$ and Q into multiple sub-embeddings $\{\hat{Z}_1, \hat{Z}_2, ...\hat{Z}_L\}$ and $\{Q_1, Q_2, ...Q_L\}$, respectively. Next, cross-attention operations are performed between each visual sub-embedding $\hat{Z}_i$ and corresponding learnable query Q_i, yielding V_i. We then concatenate $\{V_1, V_2, ...V_L\}$ to obtain the final visual input for the LLM: $V \in \mathbb{R}^{L\times D}$. Finally, the language model combines the visual embeddings V with user-provided surgical text questions S as input and performs joint reasoning to generate responses.

When training the model, we use the pre-trained model of TinyLLaVA-Video and fine-tune it with the VQA pairs of SurgPub-Video. We only keep the visual encoder frozen, fully fine-tuning the LLM and video resampler.

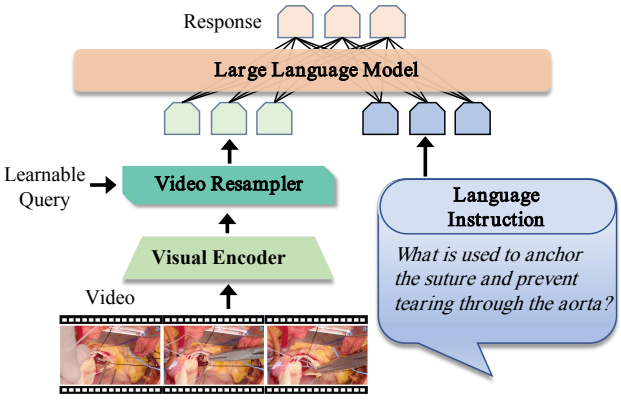


Figure 4: Architecture of SurgLLaVA-Video. A multi-frame clip passes through a video encoder, whose patch tokens are compressed by a **video resampler**; learnable queries distill visual embeddings that are concatenated with the textual prompt and decoded by the LLM to generate the answer.

Experiments

Experimental Setup

In this work, we employ GPT-4o (Hurst et al. 2024) as an agent to assist the generation of SurgPub-Video dataset. We train SurgLLaVA-Video on 4 NVIDIA A40 GPUs. During the inference, the temperature is set as 0.2. We evaluate seven different models on our constructed SurgPub-Video benchmark, including LLaVA-1.5 (Liu et al. 2023), InternVL3 (Chen et al. 2024), Qwen-2.5-VL-Instruct (Bai et al. 2025), GPT-4o (Hurst et al. 2024), Qwen2.5-Max (Bai et al. 2025), Gemini 2.0 Flash (Team et al. 2023), and our SurgLLaVA-Video. To further validate the generation ability of our SurgLLaVA-Video model, we follow the experiment setting of SurgVLM (Zeng et al. 2025) and compare SurgLLaVA-Video with seven models (six benchmarking methods and SurgVLM) on three surgical downstream tasks: surgical action recognition, surgical skill assessment, and surgical triplet recognition. In the benchmark comparison, SurgLLaVA-Video was fine-tuned on the SurgPub-Video dataset. For downstream tasks, it was jointly fine-tuned on the task training sets and SurgPub-Video. All general Multi-Modal Large Language Models (MLLMs) were evaluated using their official weights or APIs, while SurgVLM (Zeng et al. 2025) was trained with data that also includes the downstream task datasets. In particular, we use a multiple-choice question (MCQ) format that requires the model to select the correct answer from the given options and measures the accuracy by counting exact matches between predictions and ground-truth answers. In each experiment, the best result for each task is highlighted in **bold**, while the second-best result is indicated with an underline.

Model	Param.	Brst.	Card.	Endo.	Esop.	Liv.	Lung	Medi.	Orth.	Thor.	Urol.	Vasc.	Overall
LLaVA-1.5	7B	38.47	44.12	36.95	42.83	34.56	39.21	48.73	29.67	52.04	31.28	37.45	35.82
InternVL3	78B	75.00	78.72	50.00	66.67	42.86	70.00	27.27	75.00	70.79	63.16	75.18	69.17
Qwen2.5-VL-Instruct	72B	81.82	61.24	38.46	60.00	33.33	74.90	66.67	57.69	74.92	52.94	61.54	62.44
GPT-4o-2024-0806	-	81.82	67.98	50.00	66.67	57.89	58.82	53.85	78.95	69.77	77.78	62.00	66.67
Qwen 2.5 Max	-	75.00	74.19	42.86	72.00	74.03	71.53	68.18	65.66	**86.39**	67.65	73.58	73.56
Gemini 2.0 Flash	-	75.48	73.07	42.30	78.13	67.50	59.92	72.67	**81.46**	78.30	75.93	77.23	73.33
SurgLLaVA-Video (Ours)	3B	**82.14**	**84.38**	**61.38**	**89.69**	**74.19**	**82.84**	**79.76**	78.24	81.74	**80.23**	**82.73**	**82.84**

Table 2: Performance of various VLMs on proposed benchmark.

Model	Param.	RARP Action Recognition				Endoscapes CVS Balanced Accuracy			
		Accuracy	Recall	Precision	Jaccard	Average	Crit. 1	Crit. 2	Crit. 3
LLaVA-1.5	7B	23.46	14.22	11.53	6.31	49.56	48.97	50.00	49.72
InternVL3	78B	27.32	24.03	30.48	13.30	53.50	52.02	57.54	50.94
Qwen2.5-VL-Instruct	72B	28.20	13.04	22.30	5.94	47.22	50.26	41.67	49.72
GPT-4o-2024-0806	-	28.10	15.99	16.15	9.16	9.07	9.23	8.81	9.17
Qwen2.5-Max	-	28.30	14.60	10.20	7.14	45.03	44.44	44.52	46.11
Gemini 2.0 Flash	-	24.40	17.51	18.54	7.27	52.37	57.90	50.16	49.06
SurgVLM	72B	42.90	34.64	31.45	19.22	51.40	51.59	52.22	50.39
SurgLLaVA-Video (Ours)	3B	**65.65**	**45.50**	**55.35**	**33.31**	**58.88**	**60.06**	**58.11**	**62.19**

Table 3: Performance of various VLMs on RARP action recognition and Endoscapes CVS task.

Benchmark Results

We evaluate the performance of SurgLLaVA-Video and various VLMs on the proposed benchmark. Tab. 2 presents the corresponding experimental results. SurgLLaVA-Video achieves consistent and significant improvements across all evaluated organ categories and in overall performance on the benchmark. It reaches 82.84% overall accuracy, lifting performance by 16.17% over GPT-4o. Moreover, SurgLLaVA-Video achieves these advances despite having substantially fewer parameters with 3B, emphasizing that the effectiveness of incorporating the proposed SurgPub-Video dataset, which enables the model to develop advanced surgical video understanding capabilities.

Downstream Task Results

Surgical Action Recognition We use SAR-RARP dataset (Psychogyios et al. 2023) for action recognition, which annotates eight fine-grained robotic actions of prostatectomy. Model performance is measured by classification accuracy, recall, precision, and Jaccard. Tab. 3 shows the performance of each model on SAR-RARP action recognition dataset. SurgLLaVA-Video raises accuracy by more than 22% over SurgVLM, the strongest competing surgical model, while lifting precision by 23% and the Jaccard index by 14%.

Surgical Skill Assessment We benchmark on the critical view of safety (CVS) assessment using the Endoscapes2023 dataset (Mascagni et al. 2025) from laparoscopic cholecystectomy videos. We compute the widely used balanced accuracy metrics to assess performance. As shown in Tab. 3, on the Endoscapes2023 CVS task, SurgLLaVA-Video achieves an average balanced accuracy of 58.88%, surpassing the strongest baseline, InternVL3, by 5.38% and outperforming the domain-adapted SurgVLM by 7.48%. The performance improvement of SurgLLaVA-Video indicates its great potential in supporting high-risk clinical decisions. This also proves that the SurgPub-Video dataset can help the model acquire knowledge related to surgeries.

Surgical Triplet Recognition We also evaluate the model performance of triplet recognition on CholecT50 (Nwoye and Padoy 2022) dataset, which is composed of 50 surgical videos and labeled with the triplet of "tool-action-target". The Mean Average Precision(mAP) and accuracy are used as the main metrics of evaluation. Tab. 4 shows performance of each method on the CholecT50 benchmark. SurgLLaVA-Video with 3B parameters outperforms SurgVLM by 23.73% in instrument, 44.65% in verb, and 26.78% in triplet, respectively. SurgVLM keeps the top target accuracy at 57.07%, 9.95% higher than our model, yet SurgLLaVA-Video delivers a competitive target mAP of 19.93% against 21.01%.

Ablation Study

Here, we conduct ablation studies to verify the effectiveness of data scale and video length. Because different organs show consistent responses to changes in video length and data scale, whereas each task is more sensitive to these settings, we evaluate model accuracy across five tasks.

Ablation Study on Video Length Fig. 5(a) shows MCQ accuracy for the input clip length in 8, 16, 24, and 32 frames. Overall performance peaks at 16 frames with 80.82%, indicating the best trade-off between temporal context and redundancy. Instrument recognition and surgery-planning tasks follow this trend, rising from 8 to 16 frames before tapering. Anatomical structure recognition improves further to 24 frames at 85.08%, suggesting that excess frames

Model	Param.	Accuracy				mAP			
		Ins.	Verb	Target	Trip.	Ins.	Verb	Target	Trip.
LLava-1.5	7B	3.69	9.41	0.23	0.00	22.48	14.12	9.41	2.35
InternVL3	78B	38.14	9.52	3.29	0.52	22.74	14.08	9.44	2.36
Qwen2.5-VL-Instruct	72B	32.66	7.91	5.25	1.27	23.38	14.45	10.86	2.71
GPT-4o-2024-0806	-	13.33	5.89	5.94	1.50	22.80	14.34	10.42	2.60
Qwen2.5-Max	-	7.21	4.85	5.94	0.35	22.40	14.12	9.59	2.39
Gemini 2.0 Flash	-	15.18	7.44	15.87	1.85	23.14	14.45	10.91	2.54
SurgVLM	72B	62.20	18.64	**57.07**	12.52	31.44	19.16	**21.01**	6.35
SurgLLaVA-Video (ours)	3B	**85.93**	**63.29**	47.12	**39.30**	**58.98**	**42.75**	19.93	**10.61**

Table 4: Performance of various VLMs on CholecT50 triplet recognition.

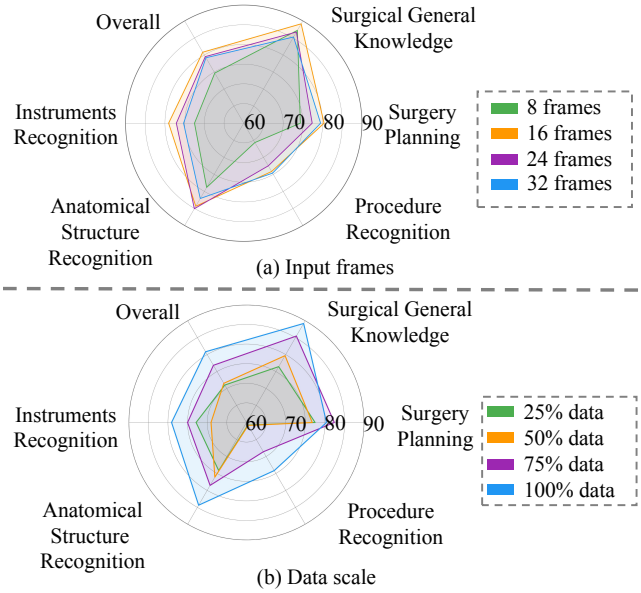


Figure 5: Ablation study of SurgLLaVA-Video. (a) Accuracy for 5 tasks and overall score under four input-frame counts. (b) Corresponding accuracy under 4 dataset scales.

add noise. Procedure recognition accuracy plateaus after 16 frames, while general knowledge scores drop beyond this length.

Ablation Study on Dataset Size Fig. 5(b) reports MCQ accuracy when the model is trained with 25%, 50%, 75%, and 100% of SurgPub-Video. Overall accuracy rises almost linearly, reaching 80.82% with the full dataset. Instrument and Procedure recognition grows by about 7%–14% from the smallest to the largest split, showing their dependence on a broad variety of visual exemplars. General-knowledge accuracy also climbs by 13%, indicating that richer textual grounding benefits from more clips. Surgery-planning peaks at 75% with 82.40% and then stabilizes, implying diminishing gains once core procedural patterns are captured. These trends confirm that most tasks continue to benefit from larger, well-balanced data, whereas planning knowledge saturates earlier.

Conclusion

In this work, we present SurgPub-Video, the first large-scale, video-based surgical dataset specifically designed to enhance VLMs' capabilities for surgical scene understanding. Addressing the existing challenges of limited data availability, semantic scarcity, and unreliable data sources, SurgPub-Video comprises more than 3,000 peer-reviewed surgical videos and more than 48,000 structured VQA pairs across multiple surgical specialties, enabling video-type input. Building upon this robust dataset, we develop SurgLLaVA-Video, a specialized and efficient surgical VLM architecture. Our model significantly outperforms general-purpose and current surgical-specific models in various benchmarks, achieving state-of-the-art performance despite using only 3 billion parameters. While SurgPub-Video and SurgLLaVA-Video demonstrate strong performance, several limitations remain. The benchmark tasks do not yet cover all organ types or surgical scenarios, and fine-tuning on specific datasets may limit generalization to unseen domains. Moreover, the current model relies on an existing architecture. Future work will explore more specialized designs for surgical video reasoning. Furthermore, we propose a comprehensive multi-task benchmark that systematically evaluates model performance across diverse surgical procedures and clinical tasks.

Acknowledgments

The work described in this paper was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project T45-401/22-N, in part by the HK RGC AoE under AoE/E-407/24-N, and in part by the Research Start-up Fund for Prof. Xiaowei Hu at the Guangzhou International Campus, South China University of Technology (Grant No. K3250310).

References

Bai, L.; Islam, M.; Seenivasan, L.; and Ren, H. 2023. Surgical-VQLA:Transformer with Gated Vision-Language Embedding for Visual Question Localized-Answering in Robotic Surgery. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 6859–6865.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Hu, Y.; Li, T.; Lu, Q.; Shao, W.; He, J.; Qiao, Y.; and Luo, P. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22170–22183.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, 1650–1654. IEEE.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, S.; Zheng, B.; Chen, W.; Peng, Z.; Yin, Z.; Shao, J.; Hu, J.; and Yuan, Y. 2025. A Comprehensive Evaluation of Multi-Modal Large Language Models for Endoscopy Analysis. *arXiv preprint arXiv:2505.23601*.

Mascagni, P.; Alapatt, D.; Murali, A.; Vardazaryan, A.; Garcia, A.; Okamoto, N.; Costamagna, G.; Mutter, D.; Marescaux, J.; Dallemagne, B.; et al. 2025. Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. *Scientific Data*, 12(1): 331.

Nwoye, C. I.; and Padoy, N. 2022. Data splits and metrics for method benchmarking on surgical action triplet datasets. *arXiv preprint arXiv:2204.05235*.

Psychogyios, D.; Colleoni, E.; Van Amsterdam, B.; Li, C.-Y.; Huang, S.-Y.; Li, Y.; Jia, F.; Zou, B.; Wang, G.; Liu, Y.; et al. 2023. Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Seenivasan, L.; Islam, M.; Kannan, G.; and Ren, H. 2023. SurgicalGPT: end-to-end language-vision GPT for visual question answering in surgery. In *International conference on medical image computing and computer-assisted intervention*, 281–290. Springer.

Seenivasan, L.; Islam, M.; Krishna, A. K.; and Ren, H. 2022. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 33–43. Springer.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. 2024. Towards generalist biomedical AI. *Nejm Ai*, 1(3): AIoa2300138.

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, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; XiXuan, S.; et al. 2024. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37: 121475–121499.

Yang, S.; Zhou, F.; Mayer, L.; Huang, F.; Chen, Y.; Wang, Y.; He, S.; Nie, Y.; Wang, X.; Sümer, Ö.; et al. 2025. Large-scale Self-supervised Video Foundation Model for Intelligent Surgery. *arXiv preprint arXiv:2506.02692*.

Yuan, K.; Kattel, M.; Lavanchy, J. L.; Navab, N.; Srivastav, V.; and Padoy, N. 2024a. Advancing surgical vqa with scene graph knowledge. *International journal of computer assisted radiology and surgery*, 19(7): 1409–1417.

Yuan, K.; Navab, N.; Padoy, N.; et al. 2024b. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *Advances in Neural Information Processing Systems*, 37: 122952–122983.

Yuan, K.; Navab, N.; Padoy, N.; et al. 2024c. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *Advances in Neural Information Processing Systems*, 37: 122952–122983.

Yuan, K.; Srivastav, V.; Yu, T.; Lavanchy, J. L.; Marescaux, J.; Mascagni, P.; Navab, N.; and Padoy, N. 2025. Learning multi-modal representations by watching hundreds of surgical video lectures. *Medical Image Analysis*, 103644.

Zeng, Z.; Zhuo, Z.; Jia, X.; Zhang, E.; Wu, J.; Zhang, J.; Wang, Y.; Low, C. H.; Jiang, J.; Zheng, Z.; et al. 2025. SurgVLM: A Large Vision-Language Model and Systematic Evaluation Benchmark for Surgical Intelligence. *arXiv preprint arXiv:2506.02555*.

Zhang, X.; Weng, X.; Yue, Y.; Fan, Z.; Wu, W.; and Huang, L. 2025. TinyLLaVA-Video: Towards Smaller LMMs for Video Understanding with Group Resampler. *arXiv preprint arXiv:2501.15513*.