
ENDOASSISTANT: A LARGE-SCALE VISION- LANGUAGE DATASET FOR ENDOSCOPIC SURGERY UNDERSTANDING FROM OPEN-SOURCE VIDEOS

Anonymous authors

Paper under double-blind review

ABSTRACT

Endoscopic interventions offer a minimally invasive approach, minimizing patient discomfort and facilitating expedited recovery. Proficient training of junior surgeons necessitates the ability to analyze and interpret endoscopic scenes through questioning and answering. Consequently, the development of a robust foundation model for endoscopic visual language understanding holds immense value for medical training and surgical education. However, existing endoscopy vision-language datasets are limited in scale and diversity, consisting of only 50 videos sourced from a few clinical sites, thus posing a significant hurdle to the advancement of generalized and robust artificial intelligence models for endoscopic surgical applications. To address this challenge, we present a large-scale, meticulously curated image-text dataset of surgical endoscopic scenes from expert surgeons, designed to propel a vision-language assistant in medical scene understanding. Encompassing 590 open-source videos spanning more than 91 hours, our curated dataset includes 65,844 unique images, 30,002 unique captions, and 157,589 image-caption/question-answering pairs. This dataset aims to assist the development of automated systems to support medical professionals by mitigating repetitive tasks. We present a comprehensive endoscopic surgery assisting pipeline, (1) a first-ever image-caption dataset specifically for endoscopic scenes; (2) an image-question-answer dataset that offers greater size and diversity compared to existing collections; (3) rigorous evaluation demonstrating its efficacy in downstream surgical endoscopic scene comprehension tasks like classification, retrieval and visual question answering.

1 INTRODUCTION

Endoscopic interventions (Sharata et al., 2013; Oblizajek et al., 2020; Bang et al., 2020) have revolutionized surgical practices by offering a minimally invasive alternative, which not only mitigates patient discomfort but also accelerates postoperative recovery periods. These procedures, facilitated by advancements in technology, entail the insertion of a slender, flexible tube equipped with a camera and light source into the body, allowing surgeons to visualize and operate within internal organs without the need for extensive incisions.

In the realm of surgical education, the proficiency of junior surgeons hinges on their ability to navigate and interpret the intricate scenes presented by endoscopic imagery. Imagine a novice surgeon, standing at the forefront of a surgical theater, peering into a monitor displaying a live feed from the endoscope inserted into a patient’s body. They must swiftly discern anatomical structures, identify abnormalities, and make critical decisions in real time, all while under the pressure of time-sensitive procedures. This demands not only technical skill but also a deep understanding of the visual cues provided by the endoscopic images. Consequently, effective training programs are highly demanded to incorporate interactive learning approaches that simulate the dynamic nature of surgical settings. By engaging in iterative cycles of questioning and answering, junior surgeons can refine their ability to analyze endoscopic scenes, anticipate challenges, and formulate strategic responses. To this end, the development of a robust foundation model for endoscopy visual language understanding emerges as a transformative endeavor in surgical education. Such a model, leveraging

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107

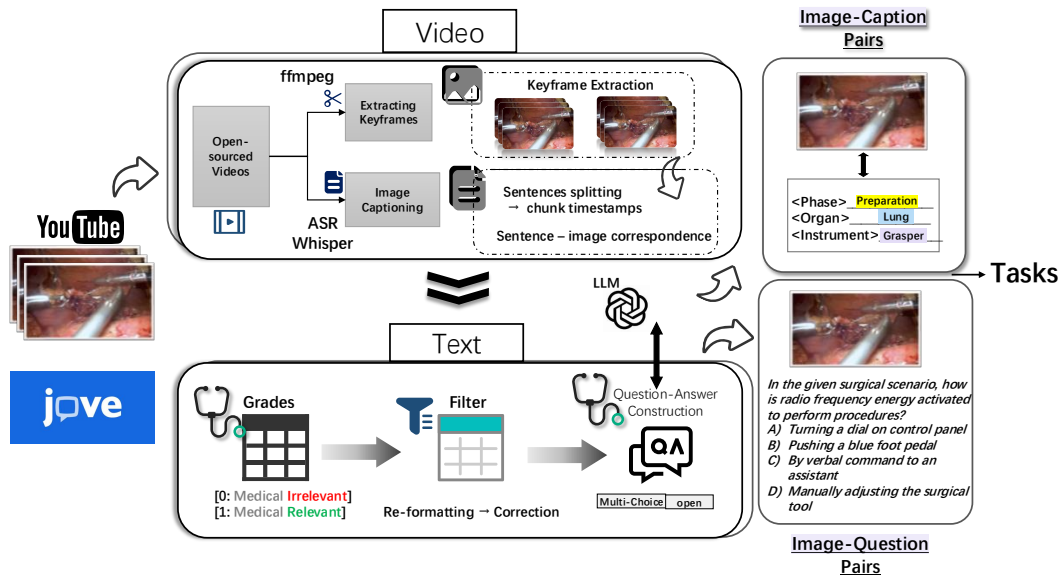


Figure 1: Working pipelines for data curation. We first collect large-scale open-sourced endoscopic surgical videos with captions from expert surgeons. The videos are chunked based on the extracted text sentences. Then keyframes are extracted and aligned with the text based on the timestamps. Both images and texts are graded and refined in collaboration with expert surgeons, supported by large vision and language models to ensure accuracy and consistency. Finally, image-caption pairs and image-question-answer pairs are generated for downstream applications including VQA, classification, and retrieval.

the power of artificial intelligence and machine learning, can systematically decode visual semantics embedded within endoscopic images.

Constructing a foundation model for endoscopic surgery understanding requires a comprehensive vision-language dataset for endoscopic surgery on a large scale. However, the current landscape of endoscopy vision-language datasets (e.g., Cholec80-VQA (Twinanda et al., 2016), PSI-AVA-VQA (Valderrama et al., 2022b)) presents a significant limitation in both scale and diversity, typically comprising fewer than 50 videos from the single clinical site. This scarcity hinges on the progress toward the creation of generalized and robust artificial intelligence models tailored specifically for endoscopic surgical applications. To address the pressing need for comprehensive datasets in the field, we introduce **EndoAssistant**, a meticulously curated large-scale collection of image-text pairs from open-source videos by expert surgeons.

Creating high-quality text-image pairs from endoscopic surgery videos presents significant challenges. Many of these videos contain background elements such as surgeon operations or procedural animations that are unsuitable for constructing our dataset. Additionally, issues like poor lighting and low image/audio quality further complicate the task. Moreover, some videos lack audio altogether or feature irrelevant content for endoscopic procedures.

To address the above issues, we meticulously designed the data collection and cleaning pipeline to ensure accurate image-text correspondence. Additionally, during the caption cleaning and question-answer construction, we collaborated with a team of expert clinical physicians to annotate and verify the accuracy and factual correctness of each pair, leveraging their domain expertise. Fig. 1 provides an overview of our dataset curation pipeline.

What sets our dataset apart is the combination of high-quality, large-scale video content and a meticulously designed data pipeline. In total, this dataset comprises 590 endoscopy videos, totaling more than 91 hours in duration, and 157,589 image-text pairs. Each image frame is accompanied by caption and question-answer correspondence designed from video captions, enhancing its utility for downstream surgical scene understanding tasks.

Another key aspect of our dataset is the detailed annotations provided by expert surgeons, which serve as invaluable repositories of human expertise in surgical procedure analysis and critical decision-making points. These captions encapsulate advanced information essential for artificial intelligence systems to not only comprehend, but also effectively replicate the nuanced decision-making processes inherent in surgical contexts. Through EndoAssistant dataset, we aim to catalyze advancements in multimodal learning and medical image understanding, empowering researchers to develop more robust and clinically applicable AI systems for endoscopic surgery.

In summary, our contributions can be outlined as follows: **i)** We meticulously curated a large-scale vision-language dataset (EndoAssistant) for endoscopic surgery comprehension, comprising 65,844 unique images, 30,002 unique captions and 157,589 image-text pairs in collaboration with expert surgeons, following rigorous data cleaning and optimization procedures. **ii)** We conduct extensive experiments using our proposed EndoAssistant (image-caption and image-question-answer pairs), demonstrating superior performance compared to previous methods on publicly available datasets. Our model utilizes contrastive learning to learn meaningful image-text corresponding representations for surgical scene understanding tasks, including cross-modal retrieval, image classification, and visual question answering.

2 RELATED WORK


Vision-Language Learning. Vision-language joint learning has been making significant strides in advanced multimodal understanding (Yin et al., 2023). A popular approach involves establishing connections between visual and textual information. Contrastive pre-training, exemplified by methods like (Wang et al., 2022a; Bao et al., 2022; Li et al., 2022; Radford et al., 2021), is one such technique. It works by simultaneously projecting text and image pairs into a shared embedding space, allowing for the discovery of relationships between the two modalities using cosine similarity. Another line of research focuses on Multimodal Large Language Models (MLLM) (Achiam et al., 2023; Team et al., 2023; Huang et al., 2023), which scale up generative pre-training (Brown et al., 2020) by incorporating both language and vision data. Early approaches like Flamingo (Alayrac et al., 2022), LLaVA (Liu et al., 2023), and MiniGPT-4 (Zhu et al., 2023) utilize pre-trained vision encoders and expand LLMs by incorporating visual tokens through visual instruction tuning. More recent advancements (Sun et al., 2023a; Lin et al., 2024; Yao et al., 2024; Wang et al., 2023; Dubey et al., 2024; Bai et al., 2023a) introduce more sophisticated joint training stages, significantly boosting performance. All these methods rely on high-quality vision-language datasets.


Vision-Language Model and Datasets for Surgical Scene Understanding. Surgical scene understanding is a complex task that involves multiple components: visual reasoning (i.e., identifying surgery phases and steps) (Seenivasan et al., 2022; Czempiel et al., 2021; Gao et al., 2021; Ding & Li, 2022; Zhang et al., 2022a), object localization (i.e., determining the position and count of organs/tools) (Qu et al., 2024; Huang et al., 2024; Jin et al., 2020; Rodrigues et al., 2022), and anomaly detection (i.e., spotting abnormalities that occur during surgery) (Jalal et al., 2023). A popular research direction is to create a unified semantic embedding that links images with their corresponding texts or captions (Twinanda et al., 2016; Zhang et al., 2022b; Vaswani et al., 2017; Yu et al., 2018).


Among all the multimodal tasks for understanding surgical scenes, the Surgical Visual Question Answering (Surgical VQA) task has been gaining increasing attention due to its interactive nature. Inspired by recent advancements in MLLM (Multimodal Large Language Models) mentioned above, several works have emerged that specifically target surgical applications. Notable methods include Surgical-VQLA (Bai et al., 2023b), which uses gated vision-language embeddings to fuse multimodal features, and Surgical-LVLM (Wang et al., 2024), which fine-tunes the MLLM Qwen-VL (Bai et al., 2023a).

However, surgical vision-language datasets are not as developed as the surgical MLLMs. The most commonly used datasets include EndoVis-18 (Allan et al., 2020), PSI-AVA (Valderrama et al., 2022a), and Cholec80 (Twinanda et al., 2016). There remains significant potential to further expand these datasets, especially in the number of categories, question diversity, and other aspects.


162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177





















 Image



 Caption

"ICG fluorescence imaging is used again to confirm the trajectory of the right and left hepatic duct before heading to the arterial and venous hydro transaction."

 Question-Answer Pairs

 	<Q> What is the next step in the procedure after introducing the guide wire through the stenosis and advancing the delivery system along it?
 	<A> Inflate the balloon to dilate the stenosis.
 	<Q> What type of lesions were primarily identified in the image?
 	<A> Tubular adenomas.
 	<Q> What issue is being observed with the surgical basket in the image?
 	<A> It is unable to adequately capture the stone.
 	<Q> What is the primary purpose of using the non-traumatic tube in this surgical scenario?
 	<A> To observe the lesion more clearly
 	<Q> How does the effectiveness of endoclip application in achieving hemostasis compare to that of thermal coagulation?
 	<A> Endoclip application is less effective than thermal coagulation

178 Figure 2: Examples of the constructed EndoAssistant dataset. Each video frame contains its corre-
179 sponding caption and question-answer pairs. We also design multiple choice questions upon the
180 captions beyond the open questions showed above.

182 3 APPROACH

184 Creating a vision-language dataset from endoscopy videos is a complex task that requires careful
185 design and refinement to extract high-quality, informative image-text correspondences from inherently
186 noisy online sources. Publicly available endoscopy videos often lack audio and suffer from low image
187 quality, making it difficult to extract scalable, expert-level data.

189 For instance, video frames containing endoscopic surgical images are frequently accompanied by
190 irrelevant visual content, such as patients, surgeons, lecturers, surgical instruments, or procedural
191 animations used for educational purposes. Text extraction poses an even greater challenge, as
192 conventional automatic speech recognition (ASR) systems often struggle to meet the specialized
193 demands of medical transcription. Both image and text de-noising add further complexity, requiring
194 surgeon verification and automation tools to ensure both the quantity and quality of the dataset. As
195 a result, using simple tools to extract frames at intervals is insufficient to capture the data accurately.
196 As a result, using simple tools to extract frames at intervals is insufficient to capture the data accurately.

197 To collect the EndoAssistant dataset, we developed custom algorithms that leverage trained models
198 to enable a semi-automated, surgeon-in-the-loop data curation pipeline, ensuring precise collection
199 and alignment of both image and text modalities for downstream applications.

200 3.1 ENDOCAPTION: EXTRACTING IMAGE AND CAPTION PAIRS

202 Our data image-text extraction pipeline comprises several key steps: 1) Retrieving video sources
203 covering the endoscopic surgery domain from Youtube ¹ and JoVE ²; 2) Filtering for videos with
204 image level pertinency; 3) Extracting and cleaning text sentences; 4) Extracting key frames; and 5)
205 Integration of both modalities.

206 **Collecting and filtering videos.** Using ‘endoscopy’ and ‘endoscopic surgery’ as retrieval keywords,
207 we collect 3372 videos from publicly available platforms, including multiple channels from YouTube
208 and medicine parts from JoVE. The diverse sources ensure a broad and representative dataset. First,
209 we used the Silero Voice Activity Detection (VAD) model ³ and only consider the video with
210 audios available. **Second, we manually evaluated the overall video quality using metrics such as**
211 **resolution, noisiness and frame rate. Videos with poor resolution or significant visual noise were**
212 **excluded to avoid introducing low-quality frames into the dataset.** To ensure the quality of relevant
213

214 ¹<https://www.youtube.com/>

215 ²<https://www.jove.com/>

³<https://github.com/snakers4/silero-vad>

216 endoscopy content, we retained only those videos among all the downloaded candidates that had more
217 than 40% of their frames classified as endoscopy images after our hierarchical image classification,
218 which leads to a total of 690 videos after filtering.

219 **Hierarchical image classification.** We used pre-trained CLIP (Radford et al., 2021) ViT-B/32
220 for coarse classification and further trained a fine classifier for endoscopic surgery images. For the
221 training data, we leveraged 62k endoscopy surgery images from EndoSLAM (Ozyoruk et al., 2021),
222 Endo-SfMLearner (Ozyoruk et al., 2021), and Kvasir (Pogorelov et al., 2017) as positive samples.
223 Additionally, we manually annotated 56k images crawled from PubMed⁴, among which 55k are
224 negative and 1k are positive. Initialized with the pre-trained Inception-v3 Szegedy et al. (2016), our
225 binary classifier is fine-tuned on the data mentioned above. This careful curation ensures high quality
226 and minimal instances of blur or poor lighting, to guarantee that the collected data meet the stringent
227 standards required for precise analysis and modeling in surgical contexts.

228 During inference, the image first goes through the pre-trained CLIP given prompts: “an endoscopic
229 or surgical image”, “a radiology image”, “a presentation slide or lecturer talking” and “Doctors
230 or patients during surgery.” This meticulously engineered the classification labels the identify and
231 differentiate endoscopic content from irrelevant content. The image with one classification probability
232 larger than 0.25 is classified as an endoscopy or non-endoscopy image. Otherwise, the image is
233 fed into the self-trained binary classifier. To evaluate this hierarchical image classification pipeline,
234 we manually curated 50k randomly sampled from all the extracted videos and got a classification
235 accuracy of 99.9% on this validation dataset.

236 **Collecting and cleaning captions.** The audio file extracted from each filtered video is fed into the
237 Whisper-large-v3 model (Radford et al., 2022) to perform Automatic Speech Recognition (ASR) and
238 transcribed into text sentences. Each sentence is treated as one chunk together with the starting and
239 ending timestamp.

240 To only retain the informative text with medical relevancy and exclude irrelevant or extraneous speech,
241 we further processed each text sentence using GPT-4 giving the below instruction.

- 242
- 243 • *You are an expert in medical terminology, specializing in analyzing and refining medical descrip-*
244 *tions. Your task is to identify and retain only the most relevant medical information from the caption,*
245 *excluding any unnecessary or irrelevant content. Correct any errors to ensure medical accuracy.*
246 *Do not introduce new information or alter the original meaning. If there is no medical relevant*
247 *information, please keep it empty.*

248 **The GPT prompt design is a surgeon-in-the-loop practice where we iteratively incorporate feedback**
249 **from a surgeon to validate the clinical relevance of the text.** To validate the efficacy of the automatic
250 text cleaning pipeline, we randomly sampled 5k from the nonempty returns and manually checked all
251 of them were **medical relevant: providing significant surgical insights (e.g., describing a procedure,**
252 **identifying anatomical features, or highlighting observations).**

253 **Extracting keyframes.** The filtered videos were partitioned into analogous chunks with identical
254 start and end times as the audio chunks. Using FFmpeg⁵, the keyframes were extracted from each
255 of the video chunks. The threshold for the software, which determines the key frame based on the
256 minimum amount of visible change in the scene, was selected after extensive experimentation. Each
257 extracted image I is accompanied by video index v and the chunk index c of the corresponding video,
258 and then goes through a hierarchical image classifier and gets a binary label y identifying if it is
259 endoscopy-related or not.

260 **Combining both modalities for correspondence.** Finally, we have a list of image-caption pairs
261 $[(v_1, c_1, T_1, \{I_1^i\}), \dots, (v_n, c_n, T_n, \{I_n^i\})]$, where T represents the text sentence and I_n^i is the of the
262 i -th images in the n -th chunk. One caption may correspond to more than one different images after
263 we apply visual content difference detection. **We note that the image-text correspondence has been**
264 **meticulously preserved throughout the entire process. While the image and text data are initially**
265 **processed through independent cleaning branches, both retain their original timestamps, providing a**
266 **strict foundation for dual-modality alignment. In the final step, these rigorous timestamps are used to**
267 **align the cleaned image and text data, ensuring both topic relevance and precise correspondence. In**

268 ⁴<https://pubmed.ncbi.nlm.nih.gov/>

269 ⁵<https://www.ffmpeg.org/>

270 addition, one surgeon manually checked 5k randomly sampled transcribed text are matched with the
271 associated time-aligned images, preserving the context of surgical actions and discussions.
272

273 3.2 ENDOQA: CONSTRUCTING IMAGE AND QUESTION-ANSWER PAIRS 274

275 From the image-caption pairs mentioned above, we further construct question-answering data that
276 might be able to facilitate clinical decision-making, enhance diagnostic precision, and beyond in
277 a more direct way. We collaborate with surgeons to design prompts that **guides GPT to generate**
278 **QA pairs from each caption text. We identified two primary issues during QA curation: (1) The QA**
279 **pairs are being used without captions, even though they are generated from caption text. (2) QA**
280 **pairs should only be used when an image is provided, and they must focus exclusively on visual**
281 **elements that can be inferred directly from the image.** After several rounds of iterative **fine-tuning**
282 **and feedback**, the prompt for question-answer generation is fixed and shown below:

- 283 • *You are an expert in endoscopic surgery. Based on the following image description, create high-*
284 *quality Q&A pairs that address both the surgical procedure and medical knowledge. Please return*
285 *high-quality Q&A, either open or multi-choice question together with the answer. The multi-choice*
286 *question includes a question, 4 options, and 1 correct answer. Please give the correct answer and*
287 *its detailed explanation.*

288 *Caption: “*****”*

289 *Q&A Guidelines:*

290 *- Questions should relate to the surgical procedure shown in the caption.*

291 *- Use proper medical terminology.*

292 *- Cover **the surgical relevant topics like actions, instruments, phases, steps, organs, objectives, and***
293 *any potential complications.*

294 *- **Assume the caption is not available and the images must exist when asking and answering.***

295 *- Provide detailed, medically accurate answers that would be useful for a medical professional.*

296 *Generate one Q&A pair that satisfies the above requirements.*

297 Our questions vary widely, encompassing topics including surgical instruments, organs, surgical
298 actions, analysis of a specific procedure and so on.

299 Fig. 2 provides qualitative examples of the curated EndoAssistant dataset: EndoCaption and EndoQA.
300

302 4 ENDOASSISTANT DATA OVERVIEW 303

304 The collected EndoAssistant, a vision-language endoscopic surgery dataset, consists of over 65K
305 unique images, 40K unique captions, and 157K associated image-text, establishing it as the largest
306 repository for endoscopic surgery vision-language tasks to date.

307 **Video Sources.** Our dataset boasts a total of 590 videos spanning over 91 hours, covering gastroin-
308 testinal endoscopy, ENT endoscopy, ophthalmic endoscopy, neuroendoscopy, and beyond. Spanning
309 a diverse range of procedures, each video within our dataset varies in duration, typically ranging from
310 2 minutes to approximately 10 minutes. The number of medical relevant text sentences per video is
311 68 on average. This extensive temporal coverage not only ensures a comprehensive representation of
312 surgical scenarios but also facilitates robust model training and evaluation across a broad spectrum of
313 procedural complexities and durations.

314 **Data Statistics.** We curated 157K endoscopy surgical image-caption pairs, associated with image-
315 QA pairs. The qualitative and quantitative analysis in Fig. 3 highlights a significant distinction between
316 our EndoQA dataset and existing repositories of endoscopic surgery VQA datasets. Specifically:
317 (1) our dataset demonstrates a more diverse semantic distribution, as evidenced by the word cloud
318 and category analysis; (2) in terms of data source, our dataset is derived from 91 hours of video and
319 65,844 relevant frames, far surpassing existing counterparts; (3) the questions in our dataset are, on
320 average, considerably longer. This length disparity is a strong indicator of the increased complexity
321 inherent in our dataset, which not only requires a deeper understanding of surgical procedures but
322 also demands more advanced reasoning capabilities from models tasked with answering these queries.
323 Consequently, our dataset presents a more challenging and invaluable resource for advancing research
in medical image understanding and multimodal reasoning.

324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377



Figure 3: A comparison between the constructed EndoAssistant and other endoscopic surgery video datasets. Our EndoAssistant excels existing datasets in terms of quality assessment (**Left**: answer category distribution and knowledge diversity) and quantity assessment (**Right**: video duration, the number of extracted frames, and the average question length).

Data Quality Samples. In order to improve the quality of image-text pairs for downstream tasks, we have explored the potential opportunities to process the text modality using context-enriched sentence expansion. To be specific, we give a brief overview of text quality with the presence of examples:

- Caption granularity.** Our endoscopic surgery dataset stands out for its inclusion of detailed captions, a distinctive feature absent in all other existing surgical video datasets. These captions are meticulously crafted by expert surgeons and encompass a wealth of knowledge, including surgical procedure analysis and background information integration. This expert knowledge serves as a crucial resource for models to emulate human expert decision-making processes effectively. To illustrate the granularity of these captions, we provide two caption examples:

Caption 1. *The following contrast-enhanced low mechanical index endosynography shows a corresponding behavior of the microvessels in autoimmune pancreatitis as already displayed for the macro vessels by contrast-enhanced high mechanical index endosynography.*

Caption 2. *Now remove the injection catheter away from the varix and the assistant begins to flush the needle continuously with distilled water.*
- Question-answering granularity.** To design high quality questions for the endoscopic surgery VQA task, we design a semi-automatic QA construction enabling the expert physicians to leverage the robust question generation capabilities of LLM. This collaborative effort yields a diverse array of questions, spanning different types and aspects of medical scenarios, thereby enhancing their applicability and challenge levels in medical education. Specifically, our designed questions cover a wide variety of endoscopic surgery aspects, including surgical phases/steps, organs, surgical, actions, and the objective of surgeons. Below are some examples.

Surgical instruments: *What is the primary advantage of using color Doppler sonography in surgical scenarios as demonstrated in the scenario where a feeding arterial vessel is observed within a tumor?*

Surgical actions: *What is the most likely action a surgical team should undertake when encountering a completely different, life-threatening scenario during surgery?*

Surgical steps: *During the procedure described, after ensuring that the snare is fully opened, what is the next step in managing the lesion according to the provided video frame caption?*

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

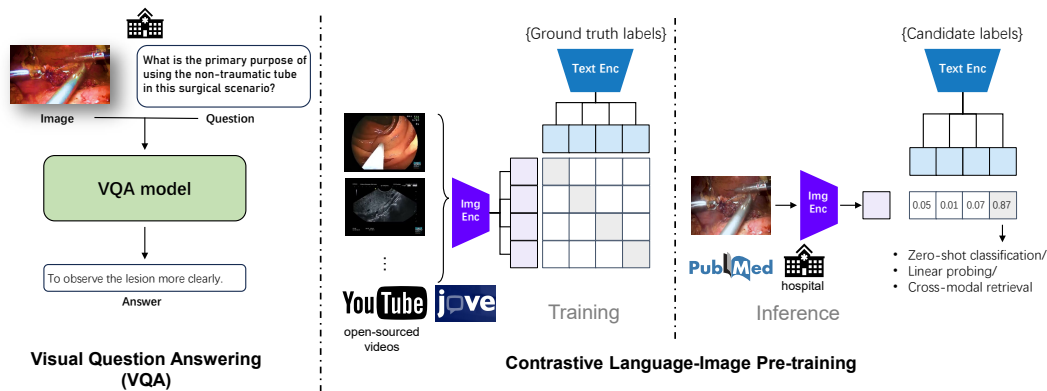


Figure 4: Illustration of the proposed **evaluation pipeline** for visual question answering (VQA) and contrastive language-image pre-training tasks. For both settings, we train the model with our curated datasets and test on publicly available datasets which are independent from our training set. This aims to demonstrate the expertise embedded in our dataset and its strong generalization ability.

Surgeon objective: *In the context of this surgical scenario, what is the primary objective of targeting biopsies to small circular glands in contrast with foci of metaplasia?*

5 EXPERIMENT

To demonstrate the effectiveness and generalizability of our proposed EndoAssistant dataset, we develop both visual question answering (VQA) models (Sec. 5.2) and contrastive language-image pre-training (CLIP) models (Sec. 5.1) using our dataset. These models are externally tested on publicly available datasets, as illustrated in Fig. 4.

For the EndoCap dataset, we apply contrastive learning techniques, following the approach in Radford et al. (2021), to unify vision and language representations within a single embedding space for endoscopic surgery. We show efficacy of EndoCap dataset through the evaluation of various external domain applications, including cross-modal retrieval and both zero-shot and few-shot image classification.

In the case of the EndoQA dataset, we train a generalist foundation vision-instruction model, designed to handle diverse VQA tasks. This model facilitates the evaluation of out-of-domain endoscopic surgery VQA datasets, thereby providing insights into the model’s ability to generalize beyond the specific domain of our training data.

5.1 CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING

In this subsection, we follow CLIP (Radford et al., 2021) to use our EndoCap dataset for contrastive language-image pre-training, where the training process simultaneously trains a text and an image encoder to increase the feature similarity from aligned pairs while decrease for misaligned pairs (Fig. 4 right).

We fine-tune CLIP with different model architectures on EndoCap and show that our data effectively enhances the image-language correspondence in the endoscopy surgery domain, consistently improving performance across multiple downstream tasks: cross-modal retrieval, zero-shot image classification, and few-shot image classification with linear probing.

Baseline models. We compare with MedCLIP (Wang et al., 2022b), PMC-CLIP (Lin et al., 2023b), BiomedCLIP (Zhang et al., 2024), CLIP (Radford et al., 2021), and SurgVLP Yuan et al. (2023; 2024a;b). Among the competing methods, MedCLIP is pre-trained on around 600K image-text pairs from the MIMIC-CXR (Johnson et al., 2019) dataset and CheXpert (Cherti et al., 2023) dataset, PMC-CLIP is pre-trained on around 642K data pairs from the PMC-OA (Lin et al., 2023b) dataset, BiomedCLIP is pre-trained on 15M data from the PMC-15M (Zhang et al., 2024) datasets, CLIP

(ViT-B/16) is pre-trained on the DataComp-1B (Gadre et al., 2023) dataset, and CLIP (ViT-B/32) is pre-trained on the LAION2B (Schuhmann et al., 2022) dataset.

Evaluation datasets. For cross-modal retrieval, we manually construct an endoscopy surgical image-caption data subset from two existing datasets: PathCap (Sun et al., 2023b) and PMC-CLIP (Lin et al., 2023b), of which the data are collected from PubMed and textbooks. For classification, four independent endoscopy surgical relevant **anatomical detection** datasets are used for external evaluation: Kvasir (Pogorelov et al., 2017), Hyper-Kvasir (Borgli et al., 2020), NBI-Inframes (Moccia et al., 2018), and GastroVision (Jha et al., 2023). **Specifically, Kvasir contains 8K images from inside the gastrointestinal (GI) tract; Hyper-Kvasir contains 11K gastrointestinal images collected during real gastro and colonoscopy examinations; NBI-Inframes contains 720 images laryngoscopy from 18 different patients affected by laryngeal spinocellular carcinoma, and GastroVision contains 8K gastrointestinal images on colon, stomach, angiectasia, and esophagitis. The categories of Kvasir include polyps, two classes related to polyp removal and three anatomical landmarks in the GI tract. The Hyper-Kvasir dataset contains 23 classes including anatomical-landmarks such as cecum, ileum, pylorus, retroflex-rectum, retroflex-stomach and z-line. In the NBI-InFrames dataset, there are four classes include tissue with intraepithelial papillary capillary loop-like vessels, leukoplakia, hypertrophic vessels, and healthy tissue. GastroVision exhibits 22 classes, including anatomical landmarks, pathological findings, polyp removal cases and beyond.**

Table 1: Cross-modal retrieval results on a cross-domain evaluation set consisting of 120 text-image pairs from PathCap (Sun et al., 2023b) and PMC-CLIP (Lin et al., 2023a) (endoscopy subsets verified by expert surgeons). Performances are measured by Recall@ (%). **Bold and underlining highlight the best and second-best performances, while shading indicates models enhanced by our curated data.**

Methods	Text-to-Image					Image-to-Text				
	R@1	R@5	R@10	R@20	R@50	R@1	R@5	R@10	R@20	R@50
MedCLIP (Wang et al., 2022b)	0.83	4.17	9.17	16.67	42.50	0.83	5.00	7.50	16.67	29.17
PMC-CLIP (Lin et al., 2023b)	<u>8.33</u>	24.17	40.83	<u>60.83</u>	83.33	5.83	<u>19.17</u>	31.67	45.83	<u>77.50</u>
BiomedCLIP (Zhang et al., 2024)	2.50	8.33	10.83	19.17	43.33	0.00	5.00	8.33	20.00	42.50
SurgVLP (Yuan et al., 2024b)	0.83	5.00	15.83	23.33	52.50	0.83	5.83	10.83	23.00	58.33
CLIP(ViT-B/32) (Radford et al., 2021)	7.50	22.50	34.17	46.67	<u>78.33</u>	3.33	11.67	20.83	37.50	64.17
CLIP(ViT-B/16) (Radford et al., 2021)	9.17	<u>26.67</u>	35.83	53.33	<u>78.33</u>	2.50	16.67	26.67	37.50	68.33
CLIP(ViT-B/32) + Ours	4.17	20.83	38.33	53.33	<u>78.33</u>	<u>4.17</u>	16.67	<u>35.00</u>	<u>49.17</u>	75.00
CLIP(ViT-B/16) + Ours	7.50	33.33	63.33	75.83	83.33	3.33	25.00	45.00	62.50	87.50

Implementation details. For contrastive language-image pretraining, we employed a batch size as 512, a learning rate as 1e-5 with 200 warm up steps. The epoch for CLIP(ViT-B/32) and CLIP(ViT-B/16) is 19 and 14 respectively. For few-shot setting, we employed a batch size as 128, MultiStepLR learning scheduler with 0.1 as learning rate and 12 epochs. For linear probing, we employed a batch size as 128, CosineAnnealingLR scheduler using a learning rate of 0.02 and weigh decay as 5e-4 using 50 epochs.

Cross-modal retrieval. We compare the cross-modal retrieval results on a subset of 120 text-image pairs extracted from PathCap and PMC-CLIP (endoscopy subsets extracted by expert surgeons) as shown in Tab. 1. Noticeably, our fine-tuned model outperforms all competing methods over most recall thresholds. The superior performance indicates its ability to better capture and align cross-modal features, resulting in more accurate and relevant retrievals.

Zero-shot image classification. We compare the zero-shot image classification results in Figure 5.a. Compared with the default CLIP pre-trained on natural images and other medical CLIP counterparts, our model fine-tuned with EndoCap significantly improves the classification accuracy, achieving the best results on the Kvasir, Hyper-Kvasir, and NBI-Inframes datasets.

Linear probing for image classification. We first use few-shot training samples with linear probing on NBI-Inframes datasets. As shown in Figure 5.b, our method consistently outperforms the counterpart methods when we increase the number of training data from 0 to 1, 4, 8, and 16. In addition, Tab. 2 compares the results leveraging all training samples with linear probing. Our fine-tuned model achieves the best results across all datasets, compared with different CLIP variants. Such performance advantages demonstrate the utility and generalization ability of our EndoCap dataset.

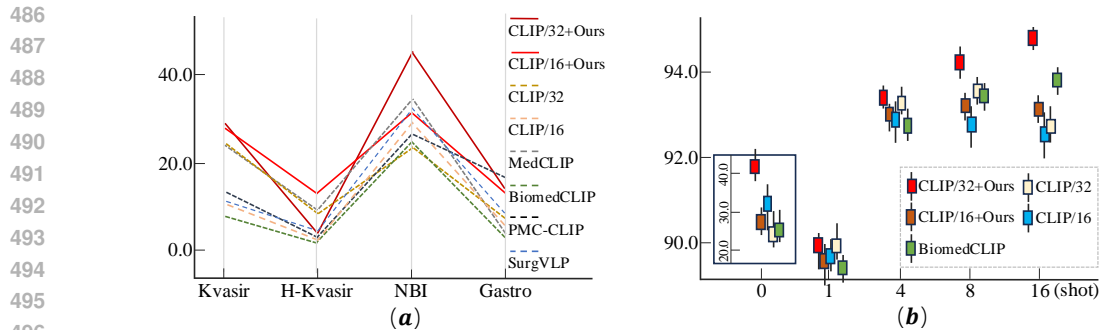


Figure 5: Comparisons of N -shot classification results. **(a)**: A zero-shot classification performance (%) comparison across Kvasir Pogorelov et al. (2017), Hyper-Kvasir Borgli et al. (2020), NBI-Inframes Moccia et al. (2018), and GastroVision Jha et al. (2023). **(b)**: The 0, 1, 4, 8, and 16-shot results on Fold1, Fold2, and Fold3 splits of the NBI-Inframes dataset Moccia et al. (2018). The box represents the mean performance (%), and the vertical line indicates the range of performance.

Table 2: Linear-probing results on Kvasir (Pogorelov et al., 2017), Hyper-Kvasir (Borgli et al., 2020), NBI-Inframes (Moccia et al., 2018), and GastroVision (Jha et al., 2023) datasets. Performances are measured by Accuracy (%). **Bold and underlining highlight the best and second-best performances, while shading indicates models enhanced by our curated data.**

Model	Kvasir	Hyper-Kvasir	NBI-Inframes			GastroVision	
			Fold1	Fold2	Fold3	Test	Val
PMC-CLIP Lin et al. (2023b)	75.69	76.37	87.08	83.75	81.67	56.05	55.56
BiomedCLIP Zhang et al. (2024)	79.64	75.12	87.08	92.50	83.75	62.63	62.41
SurgVLP Yuan et al. (2024b)	72.24	72.93	92.5	92.5	88.33	59.15	59.30
CLIP(ViT-B/32) + Ours	83.29	85.51	95.42	95.83	90.83	71.44	71.30
CLIP(ViT-B/16) + Ours	85.25	86.40	97.08	95.83	93.33	72.13	72.06

5.2 VISUAL QUESTION ANSWERING

To evaluate the efficacy of the EndoVQA dataset, we conduct visual question answering (VQA) training and evaluate on two independent datasets: EndoVis18-VQA (Allan et al., 2020) and Cholec80-VQA (Twinanda et al., 2016), which focus on two types of question formats: open and binary classification (judging the correctness of statements). We measure the performances using accuracy, recall, and Marco-F1 scores between model outputs and ground truth.

We use LLaVA (Liu et al., 2024) as our baseline and present the results in Tab. 3. **Starting with LLaVA-v1.5-7b, we finetuned with the default Lora finetuning hyper-parameters with 10 epochs.** Notably, compared with the LLaVA model trained on EndoVis18-VQA and Cholec80-VQA, finetuning it on our EndoQA improves all three metrics, enhancing overall performance on both the EndoVis18-VQA and Cholec18-VQA datasets (Tab. 3). It is important to note that MedFuse (Sharma et al., 2021) and MFB (Yu et al., 2017) utilize models specifically designed for each dataset domain, whereas our approach employs a general-purpose vision-language model for all testing datasets. As a result, our model exhibits slightly lower accuracy and recall. However, despite this, our fine-tuned LLaVA achieves the highest F1 scores across both datasets. These results underscore the endoscopy expertise and broad generalization capabilities of our dataset.

6 DISCUSSIONS

Limitations. Despite the promising results, EndoAssistant was curated using a combination of LLMs and handcrafted annotations. The automatic sections like the audio-to-text interpretation process and the subsequent text error correction can introduce biases and errors. Such errors are not rectifiable by the clinical experts in our current curation pipeline. Besides, the handcrafted sections can still benefit from automation. For example, developing a human-in-the-loop annotation pipeline using medical language models as a starting point could significantly accelerate the manual annotation

Table 3: Visual question answering results on EndoVis18-VQA (Allan et al., 2020) and Cholec80-VQA (Twinanda et al., 2016). Fine-tuning LLaVA on the proposed EndoAssistant improves the performance. **Notably, specialized models tailor different models for each specific testing task, whereas our model is general-purpose: a single model for all tasks. Bold and underlining highlight the best and second-best performances.**

Type	Methods	EndoVis18-VQA(#2769)			Cholec80-VQA(#9096)			Mean		
		Acc	Recall	F1-score	Acc	Recall	F1-score	Acc	Recall	F1-score
Specialized	MedFuse (Sharma et al., 2021)	0.6090	0.2610	0.2220	0.8610	0.3490	0.3090	0.8022	0.3285	0.2887
	MFB (Yu et al., 2017)	0.5238	0.4205	0.3622	0.8410	0.5303	0.4588	0.7669	0.5047	0.2887
	SurgicalGPT (Seenivasan et al., 2023)	0.6811	0.4649	0.4649	0.8746	0.5747	0.5794	0.8294	0.5491	0.5527
General-purpose	LLaVA (Liu et al., 2024)	0.5944	0.2988	0.2945	0.8457	0.5075	0.5001	0.7871	0.4588	0.4521
	LLaVA (Liu et al., 2024) + Ours	0.6475	0.3896	0.4008	0.8555	0.5991	0.5991	0.8069	0.5502	0.5528

process while ensuring professional accuracy at the same time. **Another area for improvement lies in correlating caption text with temporal frame sequences rather than individual frames. Currently each caption is associated with an average of 5.25 frames. In this work we simply map each frame to the caption individually. However, adopting a sequence-to-text mapping could facilitate more precise learning of visual and textual embeddings. Given that existing benchmarks primarily focus on single image-text pairs, we aim to explore training with both sequence-text and image-text data in future work. One more area for improvement is to include data with challenging conditions often faced in endoscopic surgery, such as inconsistent lighting, obstructed or blooded views. During data collection, we focused on curating high-quality samples, filtering out cases with blurriness or poor lighting, which are indeed prevalent in real-world scenarios. Moving forward, we aim to incorporate more diverse and challenging cases into our dataset to enhance scalability, specifically targeting such challenging conditions.**

Conclusion. In this work, we present an endoscopy-specialized large-scale vision-language benchmark (EndoAssist) assisting downstream tasks (cross-modal retrieval, image classification, VQA, etc.). The presented dataset is large-scale and diverse, sourced from open-source videos and verified by expert surgeons. Compared with existing endoscopic benchmarks, ours is more concentrated on surgical scenes in the operating rooms, in terms of question topics. The proposed image-caption pairs and image-question-answer pairs have demonstrated the superiority of our benchmark over mainstream vision-language pretraining frameworks through a broad range of empirical experiments.

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.
- Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023a.
- Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. 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)*, pp. 6859–6865. IEEE, 2023b.
- Ji Young Bang, Charles Melbern Wilcox, Juan Pablo Arnoletti, and Shyam Varadarajulu. Superiority of endoscopic interventions over minimally invasive surgery for infected necrotizing pancreatitis: meta-analysis of randomized trials. *Digestive Endoscopy*, 32(3):298–308, 2020.

-
- 594 Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit
595 Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-
596 of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912,
597 2022.
- 598
599 Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland,
600 Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al.
601 Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy.
602 *Scientific data*, 7(1):283, 2020.
- 603 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhari-
604 wal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agar-
605 wal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh,
606 Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler,
607 Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCand-
608 lish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot
609 learners. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan,
610 and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual
611 Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12,
612 2020, virtual*, 2020. URL [https://proceedings.neurips.cc/paper/2020/hash/
613 1457c0d6bfc4967418bfb8ac142f64a-Abstract.html](https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html).
- 614 Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade
615 Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for
616 contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer
617 Vision and Pattern Recognition*, pp. 2818–2829, 2023.
- 618 Tobias Czempiel, Magdalini Paschali, Daniel Ostler, Seong Tae Kim, Benjamin Busam, and Nassir
619 Navab. Opera: Attention-regularized transformers for surgical phase recognition. In *Medical Image
620 Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference,
621 Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pp. 604–614.
622 Springer, 2021.
- 623
624 Xinpeng Ding and Xiaomeng Li. Exploring segment-level semantics for online phase recognition
625 from surgical videos. *IEEE Transactions on Medical Imaging*, 41(11):3309–3319, 2022.
- 626
627 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
628 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
629 *arXiv preprint arXiv:2407.21783*, 2024.
- 630 Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao
631 Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, Eyal Orgad, Rahim
632 Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen
633 Musmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander
634 Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex
635 Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search
636 of the next generation of multimodal datasets, 2023.
- 637 Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. Trans-svnet: Accurate
638 phase recognition from surgical videos via hybrid embedding aggregation transformer. In *Medical
639 Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Confer-
640 ence, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, pp. 593–603.
641 Springer, 2021.
- 642
643 Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv,
644 Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning
645 perception with language models. *NeurIPS*, 2023.
- 646
647 Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu,
Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical
Image Analysis*, 92:103061, 2024.

-
- 648 Nour Aldeen Jalal, Tamer Abdulbaki Alshirbaji, Bernhard Laufer, Paul D Docherty, Thomas Neu-
649 muth, and Knut Moeller. Analysing multi-perspective patient-related data during laparoscopic
650 gynaecology procedures. *Scientific reports*, 13(1):1604, 2023.
- 651
- 652 Debesh Jha, Vanshali Sharma, Neethi Dasu, Nikhil Kumar Tomar, Steven Hicks, MK Bhuyan,
653 Pradip K Das, Michael A Riegler, Pål Halvorsen, Ulas Bagci, et al. Gastrovision: A multi-class
654 endoscopy image dataset for computer aided gastrointestinal disease detection. In *Workshop on*
655 *Machine Learning for Multimodal Healthcare Data*, pp. 125–140. Springer, 2023.
- 656
- 657 Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Multi-task
658 recurrent convolutional network with correlation loss for surgical video analysis. *Medical image*
659 *analysis*, 59:101572, 2020.
- 660
- 661 Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren,
662 Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available
663 database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- 664
- 665 Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian,
666 and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in*
667 *neural information processing systems*, 35:7290–7303, 2022.
- 668
- 669 Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On
670 pre-training for visual language models. In *CVPR*, 2024.
- 671
- 672 Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie.
673 Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International*
674 *Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536.
675 Springer, 2023a.
- 676
- 677 Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie.
678 Pmc-clip: Contrastive language-image pre-training using biomedical documents, 2023b.
- 679
- 680 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv*
681 *preprint arXiv:2304.08485*, 2023.
- 682
- 683 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in*
684 *neural information processing systems*, 36, 2024.
- 685
- 686 Sara Moccia, Gabriele O Vanone, Elena De Momi, Andrea Laborai, Luca Guastini, Giorgio Peretti,
687 and Leonardo S Mattos. Learning-based classification of informative laryngoscopic frames.
688 *Computer methods and programs in biomedicine*, 158:21–30, 2018.
- 689
- 690 Nicholas Oblizajek, Naoki Takahashi, Sevda Agayeva, Fateh Bazerbachi, Vinay Chandrasekhara,
691 Michael Levy, Andrew Storm, Todd Baron, Suresh Chari, Ferga C Gleeson, et al. Outcomes of
692 early endoscopic intervention for pancreatic necrotic collections: a matched case-control study.
693 *Gastrointestinal endoscopy*, 91(6):1303–1309, 2020.
- 694
- 695 Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L Bobrow, Gulfize Coskun, Kagan Inc-
696 etan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, et al.
697 Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach
698 for endoscopic videos. *Medical image analysis*, 71:102058, 2021.
- 699
- 700 Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas
701 de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter The-
lin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease
detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 164–169, 2017.
- 702
- 703 Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, Zongwei Zhou, et al.
704 Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks.
Advances in Neural Information Processing Systems, 36, 2024.

702 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
703 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
704 models from natural language supervision. In *International conference on machine learning*, pp.
705 8748–8763. PMLR, 2021.

706 Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever.
707 Robust speech recognition via large-scale weak supervision, 2022.

708

709 Mark Rodrigues, Michael Mayo, and Panos Patros. Surgical tool datasets for machine learning
710 research: a survey. *International Journal of Computer Vision*, 130(9):2222–2248, 2022.

711

712 Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi
713 Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski,
714 Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev.
715 Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

716

717 Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-vqa:
718 Visual question answering in surgical scenes using transformer. In *International Conference on
719 Medical Image Computing and Computer-Assisted Intervention*, pp. 33–43. Springer, 2022.

720

721 Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. Surgicalgpt: End-to-
722 end language-vision gpt for visual question answering in surgery. In *International Conference on
723 Medical Image Computing and Computer-Assisted Intervention*, pp. 281–290. Springer, 2023.

724

725 Ahmed Sharata, Ashwin A Kurian, Christy M Dunst, Neil H Bhayani, Kevin M Reavis, and Lee L
726 Swanström. Peroral endoscopic myotomy (poem) is safe and effective in the setting of prior
727 endoscopic intervention. *Journal of Gastrointestinal Surgery*, 17:1188–1192, 2013.

728

729 Dhruv Sharma, Sanjay Purushotham, and Chandan K Reddy. Medfusenet: An attention-based
730 multimodal deep learning model for visual question answering in the medical domain. *Scientific
731 Reports*, 11(1):19826, 2021.

732

733 Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao,
734 Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv
735 preprint arXiv:2307.05222*, 2023a.

736

737 Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Zhongyi Shui, Xiaoxuan Yu, Yizhi Zhao,
738 Honglin Li, Yunlong Zhang, Ruoqia Zhao, et al. Pathasst: Redefining pathology through generative
739 foundation ai assistant for pathology. *arXiv preprint arXiv:2305.15072*, 2023b.

740

741 Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking
742 the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer
743 vision and pattern recognition*, pp. 2818–2826, 2016.

744

745 Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu
746 Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable
747 multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

748

749 Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and
750 Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE
751 transactions on medical imaging*, 36(1):86–97, 2016.

752

753 Natalia Valderrama, Paola Ruiz Puentes, Isabela Hernández, Nicolás Ayobi, Mathilde Verlyck, Jessica
754 Santander, Juan Caicedo, Nicolás Fernández, and Pablo Arbeláez. Towards holistic surgical scene
755 understanding. In *International conference on medical image computing and computer-assisted
intervention*, pp. 442–452. Springer, 2022a.

756

757 Natalia Valderrama, Paola Ruiz Puentes, Isabela Hernández, Nicolás Ayobi, Mathilde Verlyck, Jessica
758 Santander, Juan Caicedo, Nicolás Fernández, and Pablo Arbeláez. Towards holistic surgical scene
759 understanding. In *International conference on medical image computing and computer-assisted
760 intervention*, pp. 442–452. Springer, 2022b.

-
- 756 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz
757 Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing*
758 *systems*, 30, 2017.
- 759 Guankun Wang, Long Bai, Wan Jun Nah, Jie Wang, Zhaoxi Zhang, Zhen Chen, Jinlin Wu, Mobarakol
760 Islam, Hongbin Liu, and Hongliang Ren. Surgical-ivlm: Learning to adapt large vision-language
761 model for grounded visual question answering in robotic surgery. *arXiv preprint arXiv:2405.10948*,
762 2024.
- 763 Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang,
764 Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv*
765 *preprint arXiv:2311.03079*, 2023.
- 766 Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal,
767 Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language:
768 Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022a.
- 769 Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from
770 unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022b.
- 771 Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li,
772 Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint*
773 *arXiv:2408.01800*, 2024.
- 774 Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on
775 multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.
- 776 Tan Yu, Jingjing Meng, and Junsong Yuan. Multi-view harmonized bilinear network for 3d object
777 recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,
778 pp. 186–194, 2018.
- 779 Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with
780 co-attention learning for visual question answering. In *Proceedings of the IEEE international*
781 *conference on computer vision*, pp. 1821–1830, 2017.
- 782 Kun Yuan, Vinkle Srivastav, Tong Yu, Joel Lavanchy, Pietro Mascagni, Nassir Navab, and Nicolas
783 Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures.
784 2023.
- 785 Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Hecvl: hierarchical video-language
786 pretraining for zero-shot surgical phase recognition. In *International Conference on Medical Image*
787 *Computing and Computer-Assisted Intervention*, pp. 306–316. Springer, 2024a.
- 788 Kun Yuan, Vinkle Srivastav, Nassir Navab, Nicolas Padoy, et al. Procedure-aware surgical video-
789 language pretraining with hierarchical knowledge augmentation. In *The Thirty-eighth Annual*
790 *Conference on Neural Information Processing Systems*, 2024b.
- 791 Bokai Zhang, Julian Abbing, Amer Ghanem, Danyal Fer, Jocelyn Barker, Rami Abukhalil, Varun Ke-
792 jriwal Goel, and Fausto Milletari. Towards accurate surgical workflow recognition with con-
793 volutional networks and transformers. *Computer Methods in Biomechanics and Biomedical*
794 *Engineering: Imaging & Visualization*, 10(4):349–356, 2022a.
- 795 Bokai Zhang, Julian Abbing, Amer Ghanem, Danyal Fer, Jocelyn Barker, Rami Abukhalil, Varun Ke-
796 jriwal Goel, and Fausto Milletari. Towards accurate surgical workflow recognition with con-
797 volutional networks and transformers. *Computer Methods in Biomechanics and Biomedical*
798 *Engineering: Imaging & Visualization*, 10(4):349–356, 2022b.
- 799 Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston,
800 Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola,
801 Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang,
802 and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen
803 million scientific image-text pairs, 2024.
- 804 Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: En-
805 hancing vision-language understanding with advanced large language models. *arXiv preprint*
806 *arXiv:2304.10592*, 2023.