Procedure-Aware Surgical Video-language Pretraining with Hierarchical Knowledge Augmentation

Anonymous Author(s)
Affiliation
Address
email

A Pretraining Dataset

A.1 Videos

We start with the videos that are used for surgical vision-language pretraining in [18]. In total, there are $1,326$ surgical lecture videos. These videos are transcribed by AWS [2] and Whisper [11] audio speech recognition (ASR) to obtain the corresponding narration texts. Furthermore, we curate the videos' metadata from the online platforms to obtain the extra keystep and abstract texts. In the phase- and video-level pretraining, we need parent- and child-level text correspondences, e.g., keystep and its corresponding narration texts, to perform procedure understanding. Therefore, we filter out the videos that do not have parent-child correspondences. In total, we have $1,007$ and 920 videos for phase- and video-level pretraining, respectively.

A.2 Misspelling Error

As the narration texts are generated from the audio using the ASR system, they usually contain many misspelling errors and fragment sentences. Therefore, we apply multiple preprocessing steps to clean the narration texts.

We first built the vocabulary based on the textbook, surgical category labels, and definition words. Specifically, we refer to the academic papers, which define the surgical phases, to curate a list of definition words and build a vocabulary that contains the words of interest. We also parse and merge the words from the textbook. In total, we obtain a vocabulary of the size of $51,640$ words. Then, we use the built vocabulary along with the spell-checking algorithm [1] to correct the misspelling errors in narration texts. The algorithm utilizes Levenshtein Distance to identify words within 2 edit distances from the original. It then cross-references these permutations (insertions, deletions, replacements, and transpositions) with a word frequency list, prioritizing words with higher occurrence frequencies as potential correct results.

B Evaluation Setup

We provide a detailed description of the downstream tasks and their settings that we apply in the experiment.

Surgical Phase Recognition. Surgical phase recognition is a proxy task to test the model's surgical scene understanding ability. It aims to classify the frame of surgical video into predefined classes (phases), requiring the model to understand the instrument and anatomy's presence and their interactions by extracting visual patterns from the surgical scene image. In this work, we ignore temporal modeling in surgical phase recognition as we focus on multi-modal representation learning.

---

[1] https://github.com/barrust/pyspellchecker/

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

Table 1: Manually designed prompts for the class names to recognize the surgical phase in Cholec80 dataset. We decompose high-level phase definitions into a few basic concepts to form the text prompts.

Phase Labels	Prompts
Preparation	In preparation phase I insert trocars to patient abdomen cavity
CalotTriangleDissection	In calot triangle dissection phase I use grasper to hold gallbladder and use hook to expose the hepatic triangle area and cystic duct and cystic artery
ClippingCutting	In clip and cut phase I use clipper to clip the cystic duct and artery then use scissor to cut them
GallbladderDissection	In dissection phase I use the hook to dissect the connective tissue between gallbladder and liver
GallbladderPacking	In packaging phase I put the gallbladder into the specimen bag
CleaningCoagulation	In clean and coagulation phase I use suction and irrigation to clear the surgical field and coagulate bleeding vessels
GallbladderRetraction	In retraction phase I grasp the specimen bag and remove it from trocar

We consider phase recognition as a frame-wise image classification problem. In the surgical phase recognition task, we evaluate the model's performance based on the publicly available datasets, including Cholec80 [15], AutoLaparo [16] and MultiBypass [7].

- **Zero-shot Evaluation.** As the surgical phase labels are high-level definitions that can be decomposed into a few basic concepts, we manually construct the contextual prompts for phase labels, as shown in Tab. 1, Tab. 2 and Tab. 3. Our constructed prompts for the class names are built with the help of clinician's comments, considering the involved surgical instruments, anatomies, and events involved in a given surgical phase.

- **Linear-probing Evaluation.** For linear-probing evaluation on the surgical phase recognition downstream datasets, we keep the visual encoder frozen and train a linear classifier on the extracted features. We do not apply any image augmentation during the training. The learning rate is scaled linearly based on the actual batch size. The model is optimized using SGD optimizer with the learning rate as 0.001 and weight decay parameter as 0.0005. We train the model for 40 epochs. We fit the model on the training and validation sets and report the performance on the separate test set. For the few-shot linear-probing evaluation, we adopt an N-way K-shot approach with a slight modification to accommodate the nature of surgical videos, which contain frames from different classes. Specifically, we select 10% of the video from the training set. This ensures that data leakage is prevented and that the number of samples per class remains similar.

Cross-modal Retrieval. Cross-modal retrieval includes text-based video retrieval and video-based text retrieval. Here, we conduct the cross-modal retrieval at three hierarchical levels. We collect 537 clip-narration (clip-level) video-text pairs, 746 phase-keystep (phase-level) video-text pairs, and 86 video-abstract (video-level) video-text pairs from hold-out testing videos of SVL [18]. There are more phase-keystep than clip-narration video-text pairs because some testing videos do not have cleaned narrations and we filter them out. For video embedding generation, we sample multiple frames from the video and average pool their image embeddings. We temporally sample 10 frames for clip-/phase-/video-level videos. We conduct the zero-shot evaluation for the cross-modal retrieval task.

C Architecture & Initialization

As mentioned before, the current surgical vision-language pretraining dataset lacks the scale necessary to pretrain a robust vision-language model from scratch, therefore a good choice of architecture and initialization is important. In this section, we conduct the experiment and study the effect of different model architectures and initializations, justifying our choice of using ResNet50 architecture with ImageNet initialization as our starting point before the video-language pretraining.

2

Table 2: Manually designed prompts for the class names to recognize the surgical phase in AutoLaparo dataset.

Phase Labels	Prompts
Preparation	I use grasper to grasp and explore the field
Dividing Ligament and Peritoneum	I divide ligament and peritoneum
Dividing Uterine Vessels and Ligament	I divide uterine vessels and ligament
Transecting the Vagina	I use the dissecting hook to transect the vagina
Specimen Removal	I remove the specimen bag and uterus
Suturing	I suture the tissue
Washing	Washing

Table 3: Manually designed prompts for the class names to recognize the surgical phase in gastric bypass dataset. We use the same prompts for both StrasBypass70 and BernBypass70. We exclude the "other" class as its definition is ambiguous.

Phase Labels	Prompts
Preparation	In preparation phase I insert trocars to the abdominal cavity and expose of the operating field
Gastric pouch creation	I cut the fat tissue and open retrogastric window at stomach
Omentum division	I grasp and lift the omentum and divide it
Gastrojejunal anastomosis	I see the proximal jejunum and determine the length of the biliary limb. I open the distal jejunum and create the gastrojejunostomy using a stapler. I reinforcement of the gastrojejunostomy with an additional suture.
Anastomosis test	I place the retractor and move the gastric tube and detect any leakage of the gastrojejunostomy
Jejunal separation	I open the mesentery to facilitate the introduction of the stapler and transect the jejunum proximal
Petersen space closure	I expose between the alimentary limb and the transverse colon and close it with sutures
Jejunojejunal anastomosis	I expose between the alimentary limb and the transverse colon and close it with sutures
Mesenteric defect closure	I expose the mesenteric defect and then close it by stitches
Cleaning and coagulation	In clean and coagulation phase I use suction and irrigation to clear the surgical field and coagulate bleeding vessels
Disassembling	I remove the instruments, retractor, ports, and camera

Backbone	Init.	Zero-shot		Linear-probing (10-shot)		Linear-probing (full-shot)	
		Cholec80	**Autolaparo**	**Cholec80**	**Autolaparo**	**Cholec80**	**Autolaparo**
ResNet50	Random	29.4 / 10.4	15.3 / 10.9	42.4 / 22.1	33.4 / 20.2	44.6 / 25.3	30.7 / 19.3
	ImageNet	34.7 / 24.4	21.3 / 16.6	55.0 / 39.9	48.5 / 32.0	63.5 / 50.3	54.3 / 41.8
	CLIP	33.8 / 19.6	18.9 / 16.2	58.9 / 42.3	45.3 / 35.3	64.9 / 55.0	53.1 / 42.1
ViT-B/16	Random	20.2 / 11.5	9.1 / 8.3	38.4 / 20.9	32.1 / 19.7	48.2 / 25.9	38.4 / 25.5
	ImageNet	42.8 / 25.1	20.5 / 15.5	57.4 / 40.5	47.8 / 31.9	60.6 / 48.9	56.3 / 44.5
	Dino	35.1 / 19.1	13.9 / 9.2	54.7 / 39.2	47.4 / 31.1	64.9 / 51.2	54.0 / 42.4

Table 4: The experiments show that the initialization largely influences the performance of surgical video-language pretraining.

- ResNet50. For ImageNet initialization, we use public IMAGENET1K_V1 weights from torchvision. Random initialization means that we random initialize the visual encoder before the hierarchical vision-language pretraining. These models' textual encoders are initialized from BioClinicalBert [6]. For CLIP initialization, we initialize the visual and textual encoder from OpenAI's weight [10].

3

71   • ViT-B/16. For ImageNet initialization, we use weights from the official Google JAX imple-
72   mentation, which is pretrained on ImageNet21k [12] and then finetune on ImageNet1k [13].
73   We use the public pretrained weights from [3] for Dino initialization.

74 In our work, we choose ResNet50 over Vision Transformer (ViT-B/16) due to its superior perfor-
75 mance and lower parameter amounts in the context of video-language pretraining for surgical data.
76 Our experiments demonstrated that ResNet50, particularly when initialized with CLIP weights,
77 outperformed ViT-B/16 across various tasks, including zero-shot and linear-probing evaluations on
78 Cholec80 and Autolaparo datasets. Despite the advanced capabilities of vision transformers, their
79 performance heavily depends on large-scale pretraining datasets, which might not always be available
80 or optimal for specialized domains like surgical scenes. Conversely, convolutional neural networks
81 like ResNet50 have shown robust generalization abilities, even when pretrained on natural images,
82 making them more suitable for our specific application. Additionally, the initialization sensitivity
83 observed in ViT-B/16 further justified our preference for ResNet50, ensuring a more reliable and
84 effective starting point for our hierarchical vision-language pretraining.

D Dynamic Time Warping

86 After achieving the cost matrix C and $\hat{C}$, we perform dynamic time warping (DTW) [14] to find
87 the minimum cost path to align the frames of video segment $V = \{v_1, ...v_T\}$ to the text sequence
88 $B = \{b_1, ...b_N\}$ and reversed text sequence $\{b_N, ...b_1\}$, respectively, as shown in Algorithm. 1. We
89 follow [17] to process the DTW function into differentiable, enabling the gradient back-propagation.
90 The differentiable loss function is the same as [5].

91 A significant advantage of using DTW is that it does not require additional temporal modules,
92 such as recurrent neural networks or attention mechanisms, to model temporal relationships. This
93 simplification allows us to focus on learning better representations by directly aligning video frames
94 and text sequences based on their semantics.

Algorithm 1 DTW to align sequences using cost matrix

1: **procedure** ALIGNSEQUENCES(C, V, B)
2:     Let T be the length of sequence V and N be the length of sequence B.
3:     Set i to T and j to N.
4:     Initialize $distance$ to 0.
5:     **while** $i > 0$ and $j > 0$ **do**
6:         $distance = distance + C[i][j]$
7:         **if** $i > 1$ and $j > 1$ and $C[i-1][j-1] \leq C[i-1][j]$ **and** $C[i-1][j-1] \leq C[i][j-1]$
    then
8:             $i \leftarrow i - 1$
9:             $j \leftarrow j - 1$
10:        **else if** $i > 1$ and $C[i-1][j] \leq C[i][j-1]$ **then**
11:            $i \leftarrow i - 1$
12:        **else**
13:            $j \leftarrow j - 1$
14:        **end if**
15:    **end while**
16:    **return** $distance$.
17: **end procedure**

E Modality Gap

96 Modality gap is a geometric phenomenon observed in the embedding space of multi-modal models [9].
97 This gap illustrates that pretrained multi-modal (vision-language) models create a joint embedding
98 space where different modalities, such as images and text, are kept at a significant distance from
99 each other. During contrastive optimization, this separation created at initialization is maintained to
100 the extent that irrelevant image embeddings can be closer to each other than to their corresponding
101 relevant text embeddings. This spatial disparity in the embedding space hinders the model's ability

4
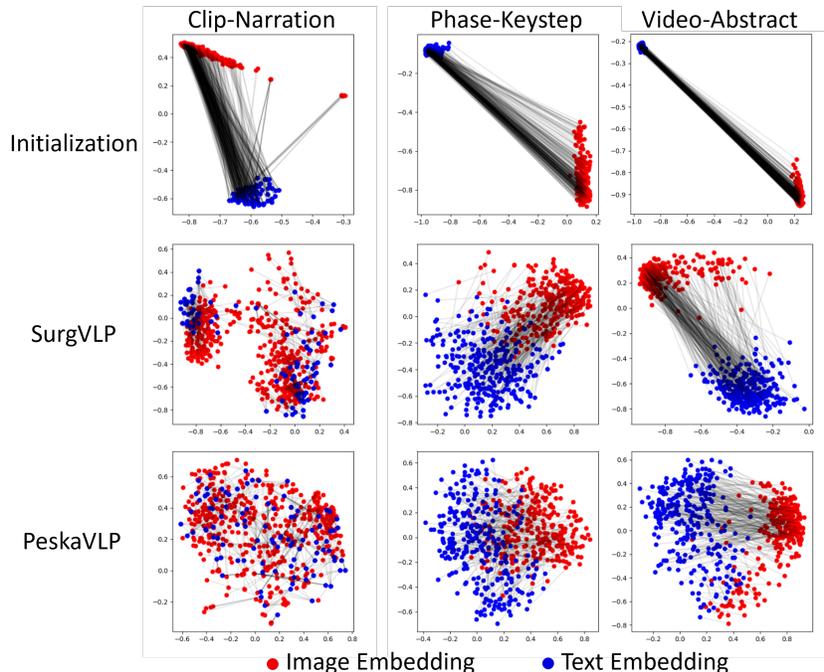
Figure 1: Modality gap visualization in different hierarchical levels. It shows that our model closes the modality gap incurred from the initialization after the hierarchical pretraining.

to effectively align and understand the relationships between visual and textual data, leading to suboptimal performance in tasks requiring integrated multi-modal comprehension. The existence of the modality gap is particularly detrimental when adapting pretrained vision-language models to cross-modal generation tasks, such as image captioning. As highlighted by several studies [8, 4], narrowing modality gap correlates with improved performance in cross-modal tasks.

As shown in Fig. 1, we visualize the embeddings of videos and their corresponding text descriptions at three hierarchical levels: clip-narration, phase-keystep, and video-abstract. Our proposed model demonstrates a significant reduction in the modality gap compared to the SurgVLP model. This alignment across different hierarchical levels ensures a more comprehensive and cohesive understanding of the multi-modal data, leading to superior performance in tasks like image captioning and other vision-language applications.

F   Surgical Phase Recognition Results

We demonstrate the zero-shot surgical phase recognition to reflect the surgical scene understanding ability of our pretrained model. Our model can identify surgical phases of different types of surgical procedures without any finetuning. Both success and failure examples are shown.

Surgical Term Understanding. In Fig. 2, we show that the pretrained model excels at identifying the "washing" phase in surgical procedures, demonstrating its capability to accurately recognize high-level surgical activities. This proficiency enhances surgical assistance systems, improving real-time analysis and decision-making in operating rooms.

Instrument Identification. In Fig. 3, we demonstrate how the visual embedding is significantly influenced by the presence of surgical instruments. Specifically, in the first row, the semantic meaning of the image changes from "calot triangle dissection" to "clip and cut" due to the appearance of a hook, even though the other anatomical features remain similar.

5
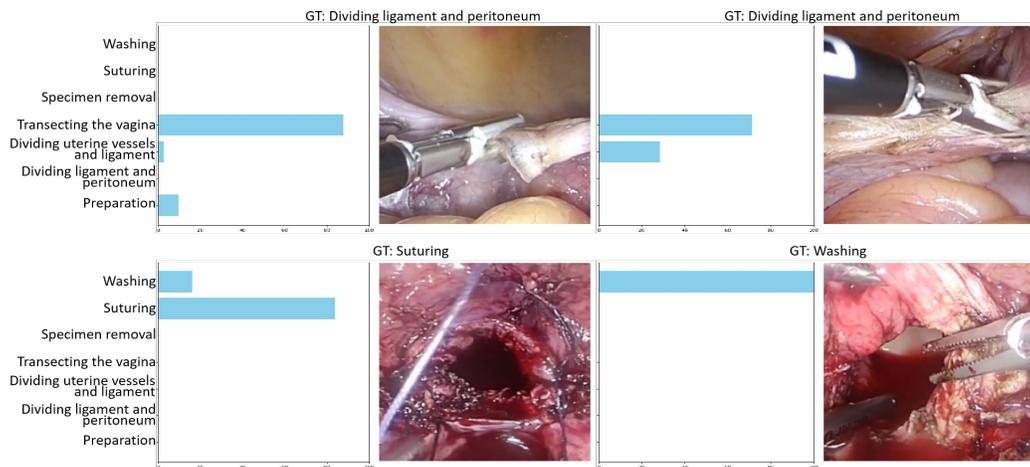
Figure 2: Qualitative surgical phase recognition results on hysterectomy. The y-axis is the class names. The x-axis is the probability of each class. The bottom right image shows that the pretrained model understands the blood fluid.
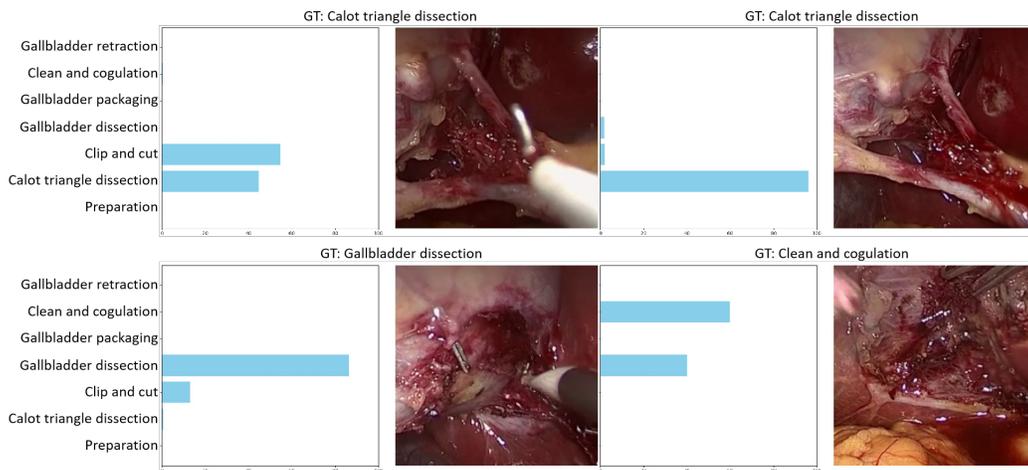
Figure 3: Qualitative surgical phase recognition results on cholecystectomy. The y-axis is the class names. The x-axis is the probability of each class. We find that the pretrained model is triggered by the instrument occurrence, such as hook in the second row.

6

G Limitations

As the pretraining process at clip-level requires additional supervision signals, i.e., visual self-supervision, the memory and computation overhead increase compared to the vanilla HecVL pretraining. Also, during the phase- and video-level pretraining, the process of dynamic time warping can be time-consuming because it is based on dynamic programming, slowing down the pretraining iteration when handling longer-term surgical videos. Additionally, the knowledge augmentation on keystep and abstract texts need to be modified to fit the other video-language pretraining datasets [1, 19] as their hierarchical paired texts are annotated manually. Instead, our knowledge augmentation is more suitable for videos in the wild from online platforms. To address these limitations, future work could focus on developing a general textual augmentation strategy using the LLM's internal knowledge, adapting to the instructional videos that miss keystep and abstract text descriptions. Furthermore, techniques for decentralizing the video-language pretraining could be explored, aiming to pretrain with multi-centric vision-language samples while preserving privacy using the federated learning strategy. This could address the scaling problem in surgical vision-language pretraining and improve the generalizationability across the centers.

H Knowledge Augmentation

Build Surgical Knowledge Base. In Fig. 4, we show that the internal surgical knowledge of large language models can be elicited to build the external knowledge base.

Build Surgical Knowledge Base. In Fig. 5, Fig. 6 and Fig. 7, we show that the knowledge of large language model can be used to enrich the semantics of the hierarchical texts, i.e., narrations, keysteps, and abstracts. Notably, it can explain high-level keystep words into descriptive sentences, enhancing textual diversity and preventing overfitting.

7

Laparoscopic right colectomy for cecal cancer
1. Position the patient appropriately for a laparoscopic right colectomy. This usually involves placing the patient in the supine position.
2. Using a trocar, create an access point to the abdomen to allow for the insertion of the laparoscope.
3. Through the laparoscope, verify the presence of the cecal cancer and its location on the right colon.
4. Identify the medial approach and begin the mobilization of the mesentery. Carefully navigate the laparoscopic tool to disconnect the right side of the colon from the rest of the organ.
5. Roam around delicately to identify the key anatomic landmarks. This could include the superior mesenteric vein, ileocolic vessels, or the duodenum.
6. Complete the division of the mesentery intracorporeally. Separate the right colon from the rest of the bowel and carefully preserve the oncologic clearance.
7. Cut the anastomosis stapler to release the healthy section of the right colon.
8. Extract the resected right colon extracorporeally through a small suprapubic incision. Take caution to make as small an incision as possible to ensure minimal harm to the patient.
9. Complete the stapled anastomosis extracorporeally. Connect the healthy section of the colon back to the rest of the organ.
10. After ensuring the anastomosis is secure and not leaking, remove the laparoscope.Please note: This is a broad outline of the steps undertaken during a laparoscopic right colectomy for cecal cancer. The specific steps may vary based on surgeon's expertise, patient's anatomy, and clinical situation.

Redo Nissen fundoplication with stapled-wedge Collis gastroplasty
1. Start the procedure by taking down the previous fundoplication.
2. Follow this by identifying the mechanism underlying the failure of the initial repair.
3. Perform an extensive mobilization of the esophagus through the hiatus to achieve an adequate length of intra-abdominal esophagus.
4. Despite the mobilization, if the esophagus remains too short, perform a Collis gastroplasty using the wedge gastrectomy technique over a 50 French bougie.
5. Following the gastroplasty, a 2.5cm of tension-free intra-abdominal esophagus should be achieved.
6. Repair the hiatus with interrupted non-absorbable sutures.
7. Finally, perform a standard Nissen fundoplication.

Stepwise approach for laparoscopic reversal of Hartmann's procedure
1. Position the patient on the operating table after administering general anesthesia to ensure patient comfort and positioning.
2. Establish pneumoperitoneum via a Veress needle to inflate the abdomen, creating a space in which to work.
3. Insert three trocars (ports) into the patient's abdomen to allow for the passage of laparoscopic instruments.
4. Inspect the abdomen with a laparoscope to locate the previous colonic stump and assess adhesions and general abdominal conditions.
5. Begin the process of adhesiolysis, involving the careful separation of adhesions between the abdominal wall and the colon.
6. Proceed with the mobilization of the colon by carefully performing a medial-to-lateral dissection.
7. Divide the colon intra-abdominally using a laparoscopic stapler, which seals off the colon and prevents leakage of bowel contents.
8. Identify the rectal stump and mobilize it within the pelvis in readiness for the reconnection of the bowel.
9. An anastomosis (connection) is created between the divided colon and the rectal stump, restoring intestinal continuity.
10. Secure the anastomosis by placing sutures and applying surgical staples to ensure a secure connection with no leakage.
11. Inspect the whole abdominal cavity visually with the laparoscope checking for any signs of bleeding, injury or any overlooked issue before ending the procedure.
12. The trocars are then removed, and the incisions sutured. The pneumoperitoneum is deflated.
13. Clean the surgical area thoroughly.
14. Dress the post-operative wounds correctly.

Laparoscopic extraction of a CBD stone after failure of ERCP (duodenal perforation)
1. The surgical area is prepared and patient is positioned for laparoscopic common bile duct (CBD) exploration.
2. Trocars are inserted at suitable locations in the abdominal region to carry out the procedure.
3. The gall bladder is reached and exposed utilizing laparoscopic tools.
4. The cystic duct is identified through careful maneuvering with laparoscopic instruments.
5. A trans-cystic approach is taken to explore the Common Bile Duct.
6. In case of large bile duct stones which cannot be extracted through the cystic duct, a choledochotomy is performed.
7. The CBD stone is visually located using the laparoscopic camera.
8. Laparoscopic instruments are used to extract the stone from the Common Bile Duct.
9. The stone is securely extracted from the body through the previously created trocar incisions.
10. Once the stone is completely removed, the common bile duct and cystic duct are checked for any potential remaining stones or blockages.
11. Procedure concludes with the removal of all laparoscopic tools and the closure of all incisions.

Figure 4: Example of surgical step knowledge base based on the large language models.

8

1. **Source:** and this be for the so be I cut the mesh just in the middle about seven centimeter link
2. **Target:** Select a mesh of appropriate dimensions that completely covers the hernia defect and extends at least 3 centimetres beyond the defect in all directions
3. **Source:** inferior epigastric vessel come from here
4. **Target:** Utilize dissection instruments to make an opening between the preperitoneal space and the transversalis fascia for easy access to the inguinal region
5. **Source:** the plain zero be often very thickened in this inflammatory condition and capsule dissection must be perform in order to help we find the plain and continued dissection
6. **Target:** Utilize meticulous dissection techniques to divide the blood vessels close to the bowel, ensuring minimal damage to the surrounding area
7. **Source:** the sigmoid colon be now or most completely release from the lateral side wall
8. **Target:** Identify and diagnose the patient with diverticulosis and chronic colo-vesical fistula
9. **Source:** we can morgue correctly define the way to proceed with the dissection
10. **Target:** Proceed with a combination of lateral and medial approach for the mobilization of the mesocolon
11. **Source:** a percutaneous suture use a straight needle be insert in the epigastric region and pass towards the apex of the right carotid
12. **Target:** Begin by positioning three ports (5mm, 12mm, 5mm) in the abdomen for laparoscopy
13. **Source:** middle colic vessel be clip and divide just above the body of the pancreas
14. **Target:** Locate the line of demarcation for the resection, ensuring to capture all the polyps and the other lesion sites observed during the preoperative investigations

Figure 5: Knowledge augmentation on the narration texts.

1. **Source:** Opening of lesser omentum
2. **Target:** The lesser omentum, a fatty apron-like structure that covers the stomach and first part of the duodenum, is opened to allow access to the stomach
3. **Source:** Start of gastric tubulization
4. **Target:** At this step, the surgeon begins creating a tube-like shape from the remaining portion of the stomach, also known as gastric tubulization
5. **Source:** End of tubulization
6. **Target:** This is when the surgeon completes the tubulization process, finalizing the smaller, sleeve-like shape of the stomach
7. **Source:** Division of greater omentum
8. **Target:** In this step, the surgeon divides the greater omentum, a large apron-like fold of visceral peritoneum that hangs down from the stomach
9. **Source:** Jejunojejunostomy
10. **Target:** The surgeon creates an opening in the two loops with a cautery hook for passage of the linear stapler and closes the opening using absorbable sutures
11. **Source:** Gastrojejunostomy
12. **Target:** The surgeon executes the gastrojejunostomy using a circular stapler, creating a connection between the stomach and jejunum
13. **Source:** Closure of Petersen's defect
14. **Target:** Towards the end, the surgeon closes Petersen's space, a potential space after Roux-en-Y gastric bypass, to prevent internal herniation
15. **Source:** Anvil placement
16. **Target:** The end of a nasogastric tube, attached to the anvil, is passed down from the mouth into the stomach
17. **Source:** Division of the ileocolic vessels
18. **Target:** The surgeon separates the blood vessels connected to the ileum and colon to prevent bleeding during the procedure
19. **Source:** Preparing the anastomosis
20. **Target:** The surgeon prepares for the anastomosis, or the surgical connection between two parts of the intestine

Figure 6: Knowledge augmentation on the keystep texts.

9

1. **Source:** This edit of a live operation demonstrates the performance of a laparoscopic gastric bypass. It demonstrates nicely manoeuvres such as retrocolic placement of the Roux limb and hand-sewn gastrojejunal anastomosis
2. **Target:** This video shows a laparoscopic gastric bypass surgery, focusing on stomach and duodenum procedures and bariatric surgery techniques for morbid obesity treatment. Main activities involve the retrocolic placement of the Roux limb and hand-sewn gastrojejunal anastomosis. They demonstrate the techniques and maneuvers used during this surgery
3. **Source:** This video shows the case of a female patient presenting with a low rectal cancer for which neoadjuvant therapy is used. The author performs a totally laparoscopic TME using a medial approach. A colorectal anastomosis without bowel protection is performed
4. **Target:** This is a surgical lecture video on a laparoscopic low anterior resection with Total Mesorectal Excision (TME) and medial mobilization of the splenic flexure in a female patient. This procedure is utilized to treat a low rectal cancer and involves the use of a medial approach. The video details how to perform a colorectal anastomosis without bowel protection. The procedure is entirely laparoscopic
5. **Source:** In this live educational video, Professor Himpens presents the case of a 34-year-old female patient (BMI of 41) with a history of morbid obesity since adolescence. She will undergo a laparoscopic sleeve gastrectomy (LSG). The preoperative work-up was normal. She had lost 2Kg six months before the procedure. Nowadays, laparoscopic sleeve gastrectomy (LSG) is one of the most commonly performed bariatric procedures. Surgical pitfalls are emphasized during the video to make sure that LSG is achieved adequately and to prevent any potential complications. In addition, trocars placement, location of the first firing of the linear stapler, the reasons why oversewing of the staple line is not performed, and thrombosis prophylaxis are also discussed during the procedure
6. **Target:** This educational video demonstrates a laparoscopic sleeve gastrectomy for a morbidly obese patient. The surgical procedure involves techniques such as the placement of trocars and the first firing of the linear stapler. It also addresses potential surgical pitfalls to ensure the adequate execution of the procedure and prevention of complications. The video highlights that oversewing of the staple line isn't performed during the procedure and also discusses the methods for thrombosis prophylaxis
7. **Source:** Intrathoracic migration of the fundoplication is one of the most common causes of failure after antireflux surgery. When the patient develops symptoms related to the volume of intramediastinal hernia, the only option is to reoperate. Such redos are complex and necessitate a thorough and painstaking approach to the potential underlying mechanisms causing intrathoracic migration, namely the length of the esophagus and cruroplasty
8. **Target:** This surgical video falls under the categories of stomach and duodenum, hiatal hernia, reflux, Nissen fundoplication, and hernia surgery. The video demonstrates a reoperation for symptomatic intrathoracic migration of a fundoplication, involving valve repositioning and reinforced crural repair. The principal activities consist of examining the underlying mechanisms causing intrathoracic migration such as the length of the esophagus and cruroplasty
9. **Source:** This video demonstrates our transumbilical three-trocar technique for single incision total colectomy and partial proctectomy with intracorporeal side-to-end ileorectal anastomosis using standard laparoscopic instrumentation. The patient is a thin 19-year-old boy with a BMI of 19 presenting with familial adenomatous polyposis (FAP). The previous colonoscopy has shown 300 polyps in the colon and very few in the distal rectum. Conventional trocars (5mm, 10mm, and 12mm) are used through a 3.5cm transumbilical incision. The ligation of the vessels is mostly carried out by the Ligasure-V vessel-sealing device using a medial-to-lateral approach. The specimen is extracted through the umbilical incision after removal of the 10mm and 12mm cannulas. The ileorectal anastomosis is carried out intracorporeally using a double stapling technique
10. **Target:** The video shows a transumbilical single incision laparoscopic total colectomy and partial proctectomy with ileorectal anastomosis performed on a 19-year-old patient with familial adenomatous polyposis. The surgery primarily uses a three-trocar technique and standard laparoscopic instruments including Ligasure-V vessel-sealing device for ligating vessels. The surgery involves making a 3.5cm transumbilical incision using 5mm, 10mm, and 12mm trocars. The colectomy specimen is extracted through the same umbilical incision. The final ileorectal anastomosis is achieved intracorporeally employing a double stapling method

Figure 7: Knowledge augmentation on the abstract texts.

10

References

[1] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23066–23078, 2023.

[2] AWS. Amazon transcribe medical, 2023.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[4] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. I can't believe there's no images! learning visual tasks using only language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2672–2683, 2023.

[5] Isma Hadji, Konstantinos G Derpanis, and Allan D Jepson. Representation learning via global temporal alignment and cycle-consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11068–11077, 2021.

[6] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[7] Joel L Lavanchy, Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Paolo Fiorini, Beat Muller-Stich, Philipp C Nett, Jacques Marescaux, Didier Mutter, and Nicolas Padoy. Challenges in multi-centric generalization: Phase and step recognition in roux-en-y gastric bypass surgery. *arXiv preprint arXiv:2312.11250*, 2023.

[8] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023.

[9] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

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

[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[12] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[14] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

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

[16] Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 486–496. Springer, 2022.

11

[17] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Kun Yuan, Vinkle Srivastav, Tong Yu, Joel Lavanchy, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. *arXiv preprint arXiv:2307.15220*, 2023.

[19] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *Proceedings of the european conference on computer vision (ECCV)*, pages 374–390, 2018.

12