# CholecMamba: A Mamba-based Multimodal Reasoning Model for Cholecystectomy Surgery

Zipei Wang[1,2], Sitian Pan[1,2], Mengjie Fang[1], Ruofan Zhang[1,2], Jie Tian[1,3(✉)][0000−0003−0498−0432], and Di Dong[1,2(✉)][0000−0003−0783−3171]

[1] Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, China
tian@ieee.org, di.dong@ia.ac.cn

**Abstract.** Automatic analysis of cholecystectomy surgical videos has significant clinical value. However, current models are limited to simple tasks like single-frame phase recognition and multi-tool classification, failing to effectively utilize video context for complex clinical reasoning. They lack the ability to integrate medical textual knowledge with cholecystectomy images and long surgical videos. We propose CholecMamba, a model that compresses video feature sequences through the Mamba architecture and deeply integrates with large-scale reasoning language models to achieve multimodal reasoning capabilities for surgical videos. Our main contributions include: 1) Designing a novel architecture that enables visual feature compression and knowledge feature injection, supporting multi-task video analysis of varying lengths; 2) Innovatively incorporating segmentation category information generated by large language models into the decoder, enhancing surgical video understanding and reasoning segmentation capabilities through medical knowledge logical reasoning; 3) Proposing the Surgical Reasoning Synthesis method, which leverages physician annotations and reinforcement learning with large language models to create the CholecReason dataset containing 49K multi-round dialogues, establishing a new benchmark for surgical video understanding and reasoning segmentation. Experimental results demonstrate that our model achieves optimal performance on existing datasets and CholecReason, with a closed-test score of 0.822, significantly outperforming the best competing model's score of 0.728. Our code is available at https://github.com/displaywz/CholecMamba.

**Keywords:** Reasoning synthesis · Surgical video analysis · Multi-class segmentation · Cross-modal learning

## 1 Introduction

Cholecystectomy is one of the most common procedures in general surgery, with its success rate and safety being critical issues in surgical practice. In recent

[1] Zipei Wang and Sitian Pan contributed equally to this work.

years, the development of computer vision and artificial intelligence has enabled computer-assisted surgical systems based on various methodologies to gradually demonstrate broad application prospects [17], [20].

Current intelligent analysis of cholecystectomy videos faces three main limitations. First, existing methods are often confined to single tasks such as surgical phase recognition or tool detection, making it difficult to balance long-sequence semantic understanding with short-term risk localization [5], [16], [3]. Second, models primarily rely on single-frame visual annotations without fully utilizing video temporal features and surgical operation semantics [12]. Finally, existing visual encoders lack capabilities in long temporal understanding and detail localization, and cannot perform surgical knowledge reasoning [14]. Therefore, there is an urgent need to develop intelligent models that can integrate long-sequence understanding with risk localization to support clinical applications.

In recent years, large language models (LLMs) and multimodal large models have achieved significant breakthroughs in natural language processing and computer vision thanks to their powerful knowledge base and reasoning capabilities [1], [21], [10]. However, directly applying these models to surgical video analysis still faces the following challenges: Current models lack deep domain-specific surgical knowledge, hindering medical reasoning and precise localization of key areas. Additionally, excessive visual token stacking limits their ability to handle varying video lengths, especially in meeting dynamic demands for long-video understanding and short-video risk localization. This underscores the anticipated challenges and the need for datasets supporting complex reasoning [2, 27].

To address these challenges, we propose CholecMamba, a multitask model framework for multimodal fusion analysis, understanding, and reasoning of surgical videos. Our motivation arises from three key issues: First, existing methods focus only on phase division and tool classification, neglecting detailed visual content [24], [8]. Second, cholecystectomy videos, with their long durations and dynamic actions, challenge current models' ability to handle varying lengths [23]. Finally, pure image or video descriptions lack clinical depth and risk model hallucinations. Medical knowledge-guided multimodal models offer the potential for solving complex medical tasks and improving clinical support.

We validated CholecMamba's effectiveness in cholecystectomy video analysis through multiple benchmark tasks. Our main research contributions include:

1. **Multitask Model Framework CholecMamba**: We propose a multi-task model framework capable of integrating multimodal data including text, images, and videos. This framework combines the text generation capabilities of LLMs with Mamba model's long sequence processing abilities, achieving semantic understanding of long videos and attention to details in short videos with logical knowledge reasoning [4], [25].
2. **Logical Reasoning Tasks for Cholecystectomy Videos**: We proposed a logical reasoning task for cholecystectomy videos and developed a reasoning segmentation assistant. For example, by asking simple questions like "What event occurred in the video?", the model can automatically segment bleeding points and report event details based on the video content.

3. **Data Annotation Method and Benchmark Dataset**: We propose a new data annotation construction method called Surgical Reasoning Synthesis. With the assistance of physicians and reinforcement learning models, we created CholecReason, a benchmark dataset for visual question-answering (VQA) and reasoning-guided segmentation in cholecystectomy videos. The dataset includes 49,404 multimodal data points, with each image or video dialogue pair containing 5-8 rounds. It incorporates rich medical visual and reasoning knowledge, filling the gap in existing datasets for multimodal tasks.

Our model achieved over 20% improvement in testing accuracy on reasoning tasks compared to Gemini-2.0-Flash, approximately 10% improvement compared to fully fine-tuned Qwen-VL-7B and LLaVA, and achieved an average Dice score of 0.973 on multi-label segmentation tasks [7], [21], [10], [19].

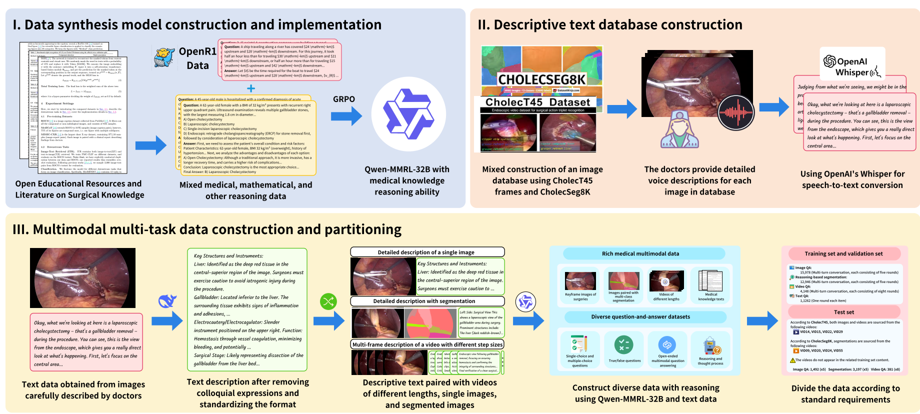


**Fig. 1.** Data construction pipeline: (1) constructing reasoning data using surgical literature and the Open-R1 dataset, (2) collecting medical professionals' voice descriptions and converting them to text, and (3) forming multi-turn reasoning data.

## 2 Methods

### 2.1 Data Construction Process

As shown in Fig. 1, we developed an accurate, efficient, and professional data construction process. Our core approach consists of two parts: accurate and efficient manual annotation, and sophisticated reinforcement learning-assisted reasoning. To achieve these conditions, we invited three surgical doctors and prepared a pre-trained Qwen-RL-32B model with reasoning capabilities for assistance.

In the early stages of data collection, we constructed reinforcement learning data primarily consisting of medical knowledge multiple-choice questions based on surgical textbooks and other standardized texts. To further enhance the model's logical reasoning abilities, we incorporated mathematical data. Using this dataset, we trained the Qwen-RL-32B model to support medical knowledge reasoning through Group Relative Policy Optimization (GRPO) [9].

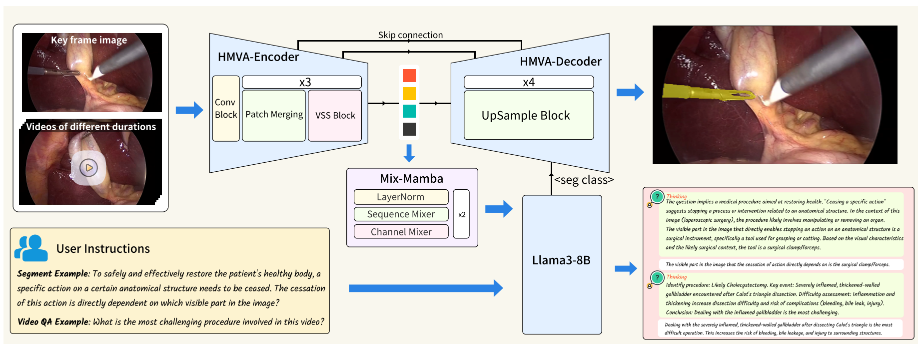


**Fig. 2.** CholecMamba architecture: An end-to-end framework for cholecystectomy video understanding and segmentation, featuring HMVA encoder-decoder, Mix-Mamba feature fusion module, and Llama3-8B based reasoning system.

The construction of the data set was based on the CholecT45 data set containing 45 surgical videos and CholecSeg8K containing 8K images [15], [6]. We first extracted frames from CholecT45 videos and combined them with CholecSeg8K to obtain 16K images. Professional doctors were invited to provide detailed audio descriptions for each image, covering both overall features and specific details. After converting the audio data to text, since the original descriptions had strong colloquial characteristics, we used LLM to clean and format the text into standardized image-text pairs. Using Qwen-RL-32B, we perform reasoning knowledge formatting on image-text pairs from gallbladder surgery video frames, generating multi-round VQA data and reasoning segmentation data, which are solely based on the textual content corresponding to the images.

To further expand the data types, we extracted key frames from the CholecT45 dataset using variable step lengths, obtaining video frame groups from different time periods (ranging from 1 to 20 minutes). These frame image descriptions were combined and input into Qwen-RL-32B to generate Video QA data.

Finally, these data were combined to construct a complete dataset. It includes 16,143 multi-round image-instruction-mask pairs, 17,470 multi-round image-QA pairs, and 4,529 multi-round video-QA pairs. Each image and segmentation data point contains 5 rounds of dialogue, while video data contains 8 rounds. The QA data includes various question types such as true/false questions, multiple-choice questions, and open-ended questions. The training and test set split follows the original partition method of CholecT45 and CholecSeg8K datasets to ensure consistency in data distribution and reliability of experimental results.

### 2.2 CholecMamba Model Architecture

**Visual Module** Through an in-depth analysis of the intrinsic relationship between dense frame processing and region segmentation, we identified a balance point between visual feature redundancy and efficient representation. Inspired by SwinUMamba, we propose a hierarchical Mamba visual architecture that efficiently processes continuous surgical video frames through an encoder-decoder pyramid network and generates precise segmentation masks for 13 anatomical structures [11]. The visual encoding process captures multiscale features from video frames through a hierarchical structure, represented as:

$$\mathbf{F}_h = \text{HMVA-Encoder}(\mathbf{I}) = \{\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_L\} \tag{1}$$

where $\mathbf{I}$ is the input surgical video frames, $\mathbf{F}_h$ are the extracted hierarchical visual features, and $\mathbf{f}_i$ is the feature map for each of the $L$ pyramid levels, capturing different spatial resolutions.

The final segmentation prediction is achieved through the decoder:

$$\mathbf{M} = \text{HMVA-Decoder}(\mathbf{F}_h, \mathbf{T}_{\text{seg}}) = \sigma\left(\sum_{i=1}^{L} \alpha_i \cdot \text{Conv}(\mathbf{f}_i \oplus \mathcal{P}(\mathbf{T}_{\text{seg}}))\right) \tag{2}$$

where $\mathbf{T}_{\text{seg}}$ is the segmentation token, $\mathbf{M}$ is the final segmentation mask, $\mathcal{P}$ is the process of associating $\mathbf{T}_{\text{seg}}$ with one of the available segmentation labels, $\oplus$ represents cross-modal feature fusion, and $\alpha_i$ are learned feature weight coefficients.

**Long Sequence Feature Fusion** As shown in Fig. 2, the designed Mix-Mamba feature aggregation module achieves efficient mapping between visual features and the LLM's semantic space through sequence mixing and channel mixing mechanisms, as well as dynamic stride feature alignment. The implementation of Mix-Mamba can be formalized as:

$$\mathbf{x}_{\text{proj}} = \text{Linear}(\mathbf{f}_L) \tag{3}$$

$$\mathbf{x}_{\text{mix}} = \mathbf{x} + \text{SeqMixer}(\text{LayerNorm}(\mathbf{x})) \tag{4}$$

$$\mathbf{V} = \mathbf{x}_{\text{mix}} + \text{ChannelMixer}(\text{LayerNorm}(\mathbf{x}_{\text{mix}})) \tag{5}$$

here, $\mathbf{x}_{\text{proj}}$ is the linear projection of feature map $\mathbf{f}_L$, with $\mathbf{x}$ being identical to $\mathbf{x}_{\text{proj}}$. $\mathbf{x}_{\text{mix}}$ captures sequence dependencies via SeqMixer, while $\mathbf{V}$ incorporates channel information through ChannelMixer. Both mixers use normalized inputs and residual connections to maintain gradient flow. The sequence mixer and channel mixer are defined as:

$$\text{SeqMixer}(\mathbf{x}) = \text{Dropout}(\text{GELU}(\mathbf{W}_s\mathbf{x})) \tag{6}$$

$$\text{ChannelMixer}(\mathbf{x}) = \text{Dropout}(\mathbf{W}_2 \cdot \text{Dropout}(\text{GELU}(\mathbf{W}_1\mathbf{x}))) \tag{7}$$

where $\mathbf{W}_s$ captures temporal dependencies between video frames, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are used for modeling feature interactions. Our cascaded Mix-Mamba projector is controlled by depth parameter $d$:

$$\text{EfficientMambaProjector}(\mathbf{x}, d) = \left(\prod_{i=1}^{d} \text{MambaMixerBlock}_i\right) \circ \text{Linear}(\mathbf{x}) \quad (8)$$

The generated visual features and instructions are jointly input to the LLM:

$$\mathbf{O}, \mathbf{T}_{\text{seg}} = \text{Llama3-8B-RL}([\mathbf{V}; \mathbf{T}_{\text{inst}}]) \quad (9)$$

Through this design, our model can efficiently process surgical video inputs at different times and achieve precise reasoning segmentation.

## 3 Experiments and Results

### 3.1 Experimental Settings

Our experiments focus on two aspects of CholecMamba: its multimodal knowledge reasoning capabilities and segmentation performance. We first conduct full training of the visual segmentation model, starting from randomly initialized weights. Next, we freeze the LLM and train Mix-Mamba using image/video-text pairs to compress and project visual features. Finally, we perform one epoch of fully supervised fine-tuning of the LLM for question-answering training. To ensure fairness, all comparison models follow the complete fine-tuning procedures provided by their respective repositories.

### 3.2 Evaluation Metrics

**Multimodal Knowledge Reasoning Capability** The model's reasoning ability is evaluated from both image and video perspectives. The tests include closed-set tests and open-ended Q&A. The inputs consist of images and randomly selected video clips of varying lengths. Closed-set questions include single-choice and multiple-choice questions. The primary evaluation metric is accuracy. Partial credit (0.5 points) is awarded for missing correct options, incorrect selections in multiple choice questions result in 0 points, and correct selections score one point. For open-ended question answering, we compute the F1 score of Bert-Score and the text embedding vector similarity (Cos (E), using OpenAI's text embedding-ada-002), evaluating the similarity between the model's final output (excluding reasoning content) and the reference answers [26].

**Reasoning-Guided Segmentation Capability** We propose a reasoning-guided segmentation task to address the current limitation of large models in accurately locating key regions. The evaluation metrics include mIoU (mean Intersection over Union), mDice (mean Dice coefficient) and F1-Score for multi-class segmentation accuracy, as well as reasoning-based annotation accuracy, which assesses whether the segmentation regions identified through reasoning align with the expected regions specified by the tasks.

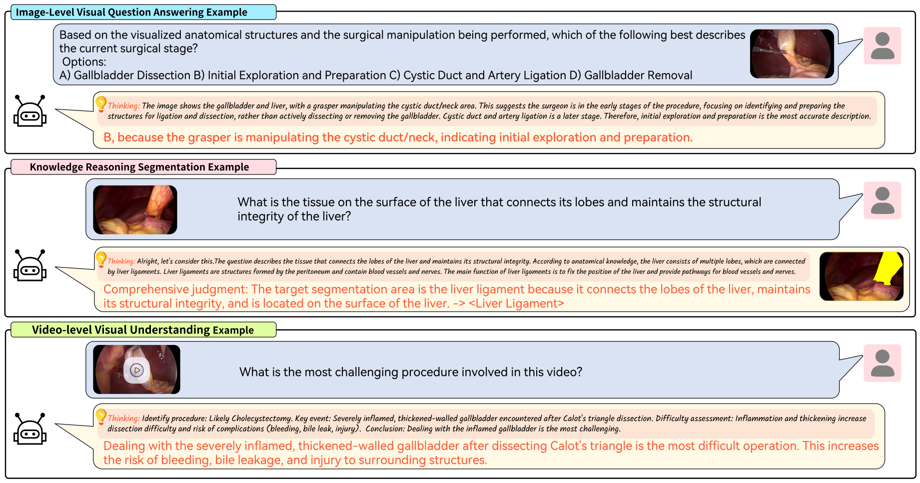


**Fig. 3.** Three Typical Application Examples of CholecMamba.

### 3.3 Results

**Multimodal Knowledge Reasoning Capability** We conducted a comparison using both closed-source and open-source models. For closed-source models, we selected the latest models with exceptional multimodal capabilities, namely Gemini-2.0-Flash and GPT-4o-1120 [19], [7]. For open-source models, we chose representative models with comparable parameter sizes, specifically Qwen2-VL-7B and LLaVA-1.6-Mistral-7B [21], [10]. We observed that directly using open-source models without fine-tuning on cholecystectomy videos did not yield testable results. Therefore, we performed full-parameter fine-tuning on the complete dataset following the procedures recommended by the respective repositories, consistent with the process used for CholecMamba. As shown in Table 1 and Fig. 3, CholecMamba demonstrates superior performance in both image and video reasoning QA tasks, particularly excelling in video QA.

**Table 1.** Performance on question answering task of different models.

| Model | VQA Closed | VQA Open | | Video Closed | Video Open | |
|---|---|---|---|---|---|---|
| | Test-score | B-score | Cos(E) | Test-score | B-score | Cos(E) |
| GPT-4o | 0.498 | 0.807 | 0.924 | 0.482 | 0.807 | 0.910 |
| Gemini | 0.613 | 0.839 | 0.934 | 0.671 | 0.752 | 0.879 |
| Qwen-VL-7B | 0.728 | 0.827 | 0.929 | 0.444 | 0.788 | 0.885 |
| LLaVA-1.6 | 0.698 | 0.829 | 0.925 | - | - | - |
| CholecMamba | **0.822** | **0.852** | **0.943** | **0.702** | **0.824** | **0.913** |

**Reasoning-Guided Segmentation Capability** As shown in Tabel 2, We performed full training on segmentation models with different architectures, including Unet, ConvNeXt-Unet, and MedSAM [18], [22], [13]. MedSAM used pre-trained weights based on 1.57M image-mask pairs and underwent complete fine-tuning for cholecystectomy segmentation. In the reasoning segmentation task, CholecMamba achieved an accuracy of 0.802, surpassing the 0.513 accuracy of Gemini-2.0-Flash and the 0.588 accuracy of Qwen2-VL-7B.

**Table 2.** Performance on segmentation of models across mDice, mIoU, and F1 score.

| Model | mDice | mIoU | F1-score |
|---|---|---|---|
| Unet | 0.9674<br>(0.9658, 0.9689) | 0.9481<br>(0.9460, 0.9500) | 0.9691<br>(0.9679, 0.9701) |
| ConvNeXt-Unet | 0.9719<br>(0.9706, 0.9734) | 0.9569<br>(0.9548, 0.9589) | 0.9656<br>(0.9645, 0.9667) |
| MedSAM | 0.9623<br>(0.9603, 0.9643) | 0.9312<br>(0.9286, 0.9338) | 0.9623<br>(0.9603, 0.9643) |
| CholecMamba | **0.9730**<br>(0.9713, 0.9749) | **0.9579**<br>(0.9557, 0.9600) | **0.9798**<br>(0.9787, 0.9808) |

### 3.4 Ablation Study

As shown in Table 3, we compared the impact of data and architecture on the model. SRS data improved performance by over 10% compared to ordinary non-inferred data. Additionally, the visual encoder and MixMamba had a significant influence on both image question answering and video reasoning.

**Table 3.** Ablation results of our method. SRS: Surgical Reasoning Synthesis method.

| Module | VQA Test-score | VideoQA Test-score |
|---|---|---|
| w/o MixMamba | 0.805 | 0.687 |
| w/o SRS | 0.712 | 0.559 |
| CLIP-ViT+Llama3-8B | 0.752 | 0.497 |
| CholecMamba (20% data) | 0.807 | 0.694 |
| CholecMamba | **0.822** | **0.702** |

## 4 Conclusion

We introduced CholecMamba for the analysis and comprehension of cholecystectomy videos, endowing the model with capabilities for processing long video sequences and reasoning. Experimental results demonstrate that CholecMamba surpasses existing models in multimodal knowledge reasoning and reasoning-guided segmentation tasks. Furthermore, our proposed SRS method and the CholecReason dataset fill the gap in available datasets for multimodal tasks in medical video analysis. This model and dataset offer a novel approach to applications in medical video analysis.

**Acknowledgments.** This work was supported by the National Key R&D Program of China (2023YFC2415200), National Natural Science Foundation of China (82441018, U24A20759, 82372053, 82361168664, 82302296), Beijing Natural Science Foundation (JQ24048), Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (CI2023C008YG), and the Science and Technology Development Fund of Macao Special Administrative Region (0006/2023/AFJ).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alabi, O., Vercauteren, T., Shi, M.: Multitask learning in minimally invasive surgical vision: A review. Medical Image Analysis p. 103480 (2025)
3. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 343–352. Springer (2020)
4. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
5. Hirsch, R., Caron, M., Cohen, R., Livne, A., Shapiro, R., Golany, T., Goldenberg, R., Freedman, D., Rivlin, E.: Self-supervised learning for endoscopic video analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 569–578. Springer (2023)
6. Hong, W.Y., Kao, C.L., Kuo, Y.H., Wang, J.R., Chang, W.L., Shih, C.S.: Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. arXiv preprint arXiv:2012.12453 (2020)
7. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
8. Jin, J., Jeong, C.W.: Surgical-llava: Toward surgical scenario understanding via large language and vision models. arXiv preprint arXiv:2410.09750 (2024)
9. Kwon, Y.D., Choo, J., Kim, B., Yoon, I., Gwon, Y., Min, S.: Pomo: Policy optimization with multiple optima for reinforcement learning. Advances in Neural Information Processing Systems **33**, 21188–21198 (2020)
10. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

11. Liu, J., Yang, H., Zhou, H.Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., et al.: Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 615–625. Springer (2024)
12. Liu, Y., Boels, M., Garcia-Peraza-Herrera, L.C., Vercauteren, T., Dasgupta, P., Granados, A., Ourselin, S.: Lovit: Long video transformer for surgical phase recognition. Medical Image Analysis **99**, 103366 (2025)
13. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
14. Nwoye, C.I., Mutter, D., Marescaux, J., Padoy, N.: Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos. International journal of computer assisted radiology and surgery **14**, 1059–1067 (2019)
15. Nwoye, C.I., Padoy, N.: Data splits and metrics for method benchmarking on surgical action triplet datasets. arXiv preprint arXiv:2204.05235 (2022)
16. Pérez, A., Rodríguez, S., Ayobi, N., Aparicio, N., Dessevres, E., Arbeláez, P.: Must: Multi-scale t ransformers for surgical phase recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 422–432. Springer (2024)
17. Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., Sestini, L., Nwoye, C.I., Hamoud, I., Sharma, S., Fleurentin, A., et al.: Dissecting self-supervised learning methods for surgical computer vision. Medical Image Analysis **88**, 102844 (2023)
18. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
19. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
20. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
21. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
22. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16133–16142 (2023)
23. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 306–316. Springer (2024)
24. Zhang, B., Sarhan, M.H., Goel, B., Petculescu, S., Ghanem, A.: Sf-tmn: Slowfast temporal modeling network for surgical phase recognition. International Journal of Computer Assisted Radiology and Surgery **19**(5), 871–880 (2024)
25. Zhang, H., Zhu, Y., Wang, D., Zhang, L., Chen, T., Wang, Z., Ye, Z.: A survey on visual mamba. Applied Sciences **14**(13), 5683 (2024)
26. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)

27. Zhu, N., Zhang, N., Shao, Q., Cheng, K., Wu, H.: Openai's gpt-4o in surgical oncology: Revolutionary advances in generative artificial intelligence. European journal of cancer **206** (2024)