MICCAI

Surgical-MambaLLM: Mamba2-enhanced Multimodal Large Language Model for VQLA in Robotic Surgery

Pengfei Hao[1], Hongqiu Wang[1], Shuaibo Li[1], Zhaohu Xing[1], Guang Yang[2], Kaishun Wu[1], and Lei Zhu[1,3(✉)]

[1] Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
`leizhu@ust.hk`
[2] Imperial College London
[3] The Hong Kong University of Science and Technology

Abstract. In recent years, Visual Question Localized-Answering in robotic surgery (Surgical-VQLA) has gained significant attention for its potential to assist medical students and junior doctors in understanding surgical scenes. Recently, the rapid development of Large Language Models (LLMs) has provided more promising solutions for this task. However, current methods struggle to establish complex dependencies between text and visual details, and have difficulty perceiving the spatial information of surgical scenes. To address these challenges, we propose a novel method, Surgical-MambaLLM, which is the first to combine Mamba2 with LLM in the surgical domain, that leverages Mamba2's ability to effectively capture cross-modal dependencies and perceive spatial information in surgical scenes, thereby enhancing the LLMs' understanding of surgical images. Specifically, we propose the Cross-modal Bidirectional Mamba2 Integration (CBMI) module to leverage Mamba2 for effective multimodal fusion, with its cross-modal integration capabilities. Additionally, tailored to the geometric characteristics of surgical scenes, we design the Surgical Instrument Perception (SIP) scanning mode for Mamba2 to scan the surgical images, enhancing the model's spatial understanding of the surgical scene. Extensive experiments demonstrate that our Surgical-MambaLLM model outperforms the state-of-the-art methods on the EndoVis17-VQLA and EndoVis18-VQLA datasets, significantly improving the performance of the Surgical-VQLA task.

Keywords: Robotic-assisted surgery · Multimodal Large Language Model · Mamba · Surgical visual question localized-answering

1 Introduction

Intelligent robotic surgery systems and the task of Visual Question Localized-Answering in robotic surgery (Surgical-VQLA) have garnered significant attention in recent years [4]. Surgical-VQLA can answer questions about organs, the location of surgical instruments, and surgical procedures based on surgical images, while simultaneously providing the corresponding bounding boxes. This

task not only assists medical students and junior residents in resolving questions about surgical procedures, thereby reducing the workload of experienced surgeons [26], but also enhances the interpretative capabilities of intelligent surgical robotic systems in understanding surgical scenes [4,30,29].

Current Surgical-VQLA methods [4,3,5,12] have achieved significant success in performance by using pre-trained vision and language models as backbone networks. In recent years, Large Language Models (LLMs) have made remarkable progress in natural language processing (NLP) [21]. Inspired by this, [8] first introduced LLMs into the surgical domain and achieved remarkable results, illustrating the immense potential of LLMs in this field. However, these methods still have some limitations. Firstly, current methods primarily rely on the Transformer-based method [28,18] for cross-modal fusion, which causes the models to focus more on global features while neglecting local details. Thus, it becomes challenging to capture visual details and establish dependencies with the text, which are essential for accurately answering questions related to the state of the instruments. Additionally, although some research [8,12,11] has introduced LLMs to surgery-related tasks, LLMs still face significant challenges in understanding surgical scenes, particularly in perceiving spatial information due to the complexity of laparoscopic environments.

Recently, the Mamba model, leveraging its State Space Models (SSMs) [16], has demonstrated the ability to efficiently capture complex dependencies within sequences while preserving detailed information. The original Mamba model performs unidirectional scanning of text sequences, which is not suitable for spatially-aware visual tasks [35]. However, some methods [35,32,31] have proved that multi-directional scanning can enhance Mamba's spatial perception capabilities for 2D vision. Inspired by this, we explore the possibility of combining Mamba with LLMs to leverage Mamba's ability to effectively capture cross-modal dependencies and perceive spatial information in surgical scenes, thereby enhancing LLMs' understanding of surgical images.

In this work, we propose the Surgical-MambaLLM model, the first method in the surgical domain to combine the Mamba2 [9] model with LLM to address Surgical-VQLA. To achieve effective cross-modal fusion and accurately perceive spatial information in surgical scenes, we design the Cross-modal Bidirectional Mamba2 Integration (CBMI) module. This module facilitates the establishment of complex dependencies between visual details and questions. Within CBMI, the Surgical Instrument Perception (SIP) scanning mode is designed instead of unidirectional sequence scanning [10], which can enhance Mamba2's spatial awareness and understanding of surgical images. Subsequently, the fused features are processed through a projector, concatenated with the textual features, and then input into the LLM. To improve the performance of the LLM, we fine-tune it using the LoRA [14] technique, thereby obtaining the final feature representation. Finally, the output features are fed into the answer prediction head and the position prediction head to obtain the final prediction results.

Overall, contributions are as follows: (1) We propose the Surgical-MambaLLM, which is the first method that integrates Mamba2 with Large Language Model
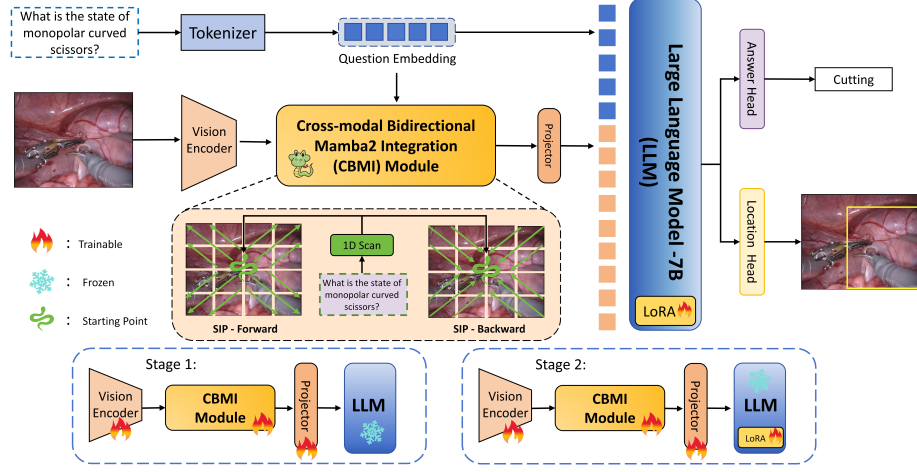
Fig. 1. An overview of the framework of our Surgical-MambaLLM. Questions are input into the tokenizer to obtain the question embedding, while surgical images are processed by the vision encoder to extract the visual features. These features are integrated within the CBMI module, which utilizes our SIP scanning mode to scan the vision features and employs modified bidirectional Mamba2 blocks for multimodal feature fusion. The fused features are then projected into the LLM to generate answer and location predictions. The training process involves two stages: initially training the vision encoder, CBMI, and projector with frozen LLM parameters, followed by fine-tuning LLM using LoRA.

in the surgical domain. By leveraging Mamba2's ability to capture complex dependencies between different modalities, we achieve effective multimodal fusion, significantly enhancing LLM performance in Surgical-VQLA tasks. (2) We introduce the Cross-modal Bidirectional Mamba2 Integration (CBMI) module, which explores effective methods for fusing visual and textual data within Mamba2. (3) We design an innovative Surgical Instrument Perception (SIP) scanning mode in CBMI to enhance Mamba2's spatial understanding of surgical images. (4) We conduct extensive experiments to validate the effectiveness of Surgical-MambaLLM in the Surgical-VQLA task. The experimental results demonstrate that our approach outperforms other State-Of-The-Art (SOTA) models on publicly available EndoVis17-VQLA and EndoVis18-VQLA datasets.

2   Methodology

We propose the Surgical-MambaLLM model, a novel method integrating Mamba2 with LLM to improve LLM's understanding of surgical scenes. The overall architecture of our Surgical-MambaLLM is illustrated in Fig. 1.
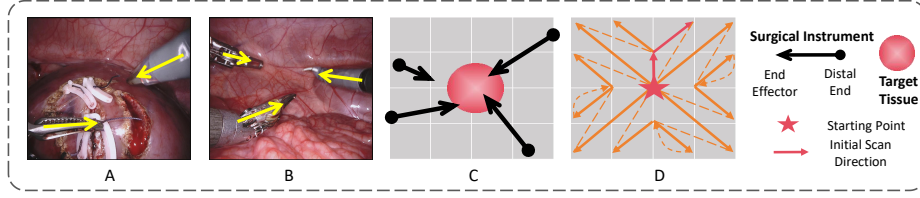
Fig. 2. A and B illustrate the directions of surgical instrument operations in surgical images; C represents the geometric modeling of the surgical scene; D is the Surgical Instrument Perception (SIP) scanning mode we proposed.

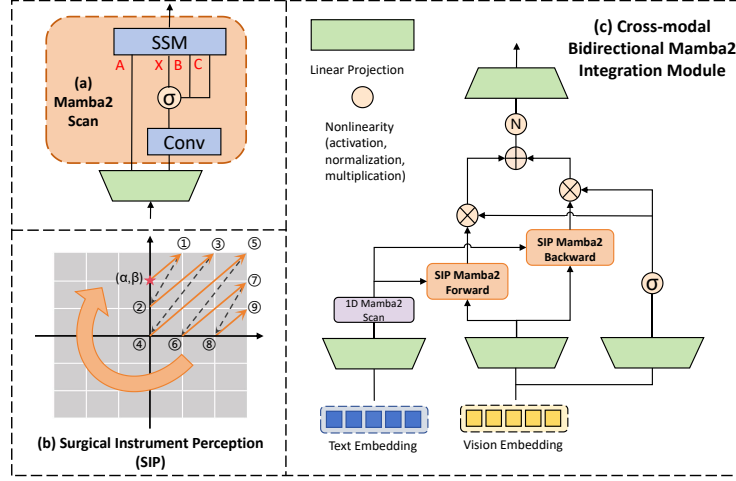### 2.1 Cross-modal Bidirectional Mamba2 Integration Module

Surgical Instrument Perception Scanning Mode In robotic surgery, surgeons typically operate from a console, observing the procedure on screens and manipulating control sticks to guide the robotic instruments [22]. To facilitate the surgeon's operation, the target organ or tissue is usually in the center of the laparoscopic image, while the ends of the robotic instruments extend from the periphery towards the central target organ [17], as illustrated in Fig. 2 A and Fig. 2 B. This observation indicates that the operation of surgical instruments in laparoscopic surgery exhibits a clear directionality from the periphery towards the center, which can be geometrically abstracted as a radial pattern from the center outward, as shown in Fig. 2 C. Therefore, we propose the Surgical Instrument Perception (SIP) scanning mode for the Mamba2 model, which performs a radial scan from the center towards four directions, ultimately scanning the entire image to obtain a global representation, as shown in Fig. 2 D. By scanning from the center outward, the SIP scanning mode is highly likely to capture continuous regions of instrument features within the image, rather than dispersing the instrument's image features across different scanning areas. This approach helps maintain the integrity of the instrument regions, thereby enhancing the mamba2's spatial awareness of the surgical scene.

Specifically, taking the first quadrant as an example, the scanning starts at (α, β), as shown in Fig. 3 (b). The maximal value of x_n or y_n is N, and the trajectory can be described by the following formula:

$$(x_{n+1}, y_{n+1}) = \begin{cases} (0, y_n - k_n) & \text{if } y_n = N, x_n \neq N \\ (x_n - k_n, 0) & \text{if } x_n = N \\ (x_n + 1, y_n + 1) \end{cases}, \qquad (1)$$

$$k_n = \begin{cases} x_n + 1 & \text{if } y_n > x_n \\ y_n - 1 & \text{if } y_n \leq x_n \end{cases}, \qquad (2)$$

where (x_n, y_n) is the current point, $(x_n + 1, y_n + 1)$ is the next point and the initial point is $(x_0, y_0) = (\alpha, \beta)$. k_n is a dependent variable that can be denoted as (2). Subsequently, we sequentially scan the fourth, third, and second quadrants, ultimately obtaining the features of the entire image.

Fig. 3. (a) is the architecture of Mamba2 Scan; (b) presents the details of the scanning track; (c) is the framework of the CBMI module.

Cross-modal Bidirectional Mamba2 In the multimodal domain, most existing Mamba methods [23,34] typically only concatenate visual features with textual features without designing fusion methods specifically for the two modalities. Therefore, we propose the Cross-modal Bidirectional Mamba2, which performs bidirectional scanning of visual features and textual features through the SIP scanning mode to achieve efficient feature fusion and improve the model's spatial understanding of surgical scenes, as shown in Fig. 3 (c).

Specifically, first, the text embedding t and the visual embedding v are preliminarily processed through linear projections to get visual feature F_t and text feature F_v:

$$F_t = l_t(t), F_v = l_v(v), \tag{3}$$

where l_t and l_v are linear layer for feature projection.

Subsequently, F_v and F_t are input into the SIP-Mamba2 Forward and SIP-Mamba2 Backward module. In this module, F_t undergoes 1D scanning, and F_v undergoes bidirectional scanning in the SIP scanning mode to obtain sequence features in opposite directions. Then F_t and the bidirectional F_v are input into the Mamba2 Scan to obtain the fused forward sequence features S_{forward} and backward sequence features S_{backward}. The structure of the Mamba2 Scan is shown in Fig. 3 (a). The formula is as follows:

$$S_{\text{forward}} = \text{SIP-Mamba2}_{\text{forward}}(F_t, F_v), \tag{4}$$

$$S_{\text{backward}} = \text{SIP-Mamba2}_{\text{backward}}(F_t, F_v), \tag{5}$$

where SIP-Mamba2$_{\text{forward}}$ is SIP-Mamba2 Forward module, and SIP-Mamba2$_{\text{backward}}$ is SIP-Mamba2 Backward module. Next, S_{forward} and S_{backward} are multiplied

by the original features F_v activated by σ and then added together to obtain the fused features S. Finally, normalization and linear projection are applied to obtain the output features S_{output}. The formula is as follows:

$$S = S_{\text{forward}} \cdot \sigma(F_v) + S_{\text{backward}} \cdot \sigma(F_v), \tag{6}$$

$$S_{\text{output}} = Linear(LN(S)), \tag{7}$$

where σ is the activation function, $Linear$ is a linear layer and LN denotes Layer normalization.

2.2   Training Strategy

Our training strategy consists of two stages for efficient model training. In the first stage, we freeze the parameters of LLM and train the vision encoder, CBMI module, projector, answer head and location head. This stage aims to enable the CBMI module to fuse visual and textual features effectively and adapt its output to the LLM via the projector. This process ensures that the vision encoder and the CBMI module can fully understand and process the complex information present in surgical scenarios. In the second stage, we fine-tune the LLM using LoRA techniques, applying a lower learning rate. This fine-tuning further optimizes the LLM for surgical-VQLA tasks while preserving the initial learning outcomes.

3   Experiments and Results

3.1   Dataset

EndoVis-18-VQLA This Dataset from the 2018 MICCAI Endoscopic Vision Challenge [1] includes video sequences from 14 surgeries. The annotations for EndoVis-18-VQLA are accessible via [4]. Recent work [5] has expanded the training set from 9014 to 12741 QA pairs and the test set from 2769 to 3820 QA pairs. Following [5], we used the expanded dataset to ensure robustness and credibility in our study. Following previous works [4,3,5], it includes 1,560 training frames and 436 test frames.
EndoVis-17-VQLA This Dataset from the 2017 MICCAI Endoscopic Vision Challenge [2] features 10 robotic surgery video sequences. Annotations are available publicly [4]. Following guidelines in [4,3,5], we use this dataset for external validation to test our model's generalization on unseen data. The dataset has been expanded from 472 to 708 QA pairs [5], enhancing our study's robustness, credibility, and fairness. Models' performance will be evaluated on these new, unseen surgical scenarios.

3.2   Implementation Details

We employ the CLIP-ViT-B/32 [24] pre-trained model as our vision encoder. The backbone of the CBMI module is Mamba2-130M [9], which comprises 24 layers.

Table 1. Comparison experiments between our Surgical-MambaLLM and other methods on EndoVis-18 and EndoVis-17 datasets

Models	EndoVis - 18			EndoVis - 17		
	Acc	F-Score	mIoU	Acc	F-Score	mIoU
VisualBERT [26]	0.6234	0.3269	0.7336	0.4516	0.2698	0.7268
VisualBERT RM [26] (MICCAI'22)	0.6365	0.3087	0.7463	0.4622	0.2865	0.7331
MFH [33]	0.5942	0.3273	0.7541	0.4614	0.3326	0.7237
BlockTucker [7]	0.6268	0.2964	0.7631	0.4552	0.3122	0.7612
MUTAN [6]	0.6298	0.3379	0.7714	0.4784	0.3244	0.7694
GVLE-LViT [4] (ICRA'23)	0.6512	0.3365	0.7739	0.4565	0.2679	0.7296
CAT-ViL DeiT [3] (MICCAI'23)	0.6436	0.3421	0.7712	0.4765	0.3467	0.7621
Surgical-VQLA++ [5] (INFORM FUSION'25)	0.6573	0.3203	0.7956	0.4983	0.4365	**0.7764**
EnVR-LPKG [12](JBHI'25)	0.6723	0.3826	0.7894	0.4786	0.4126	0.7438
Surgical-MambaLLM (our)	**0.6964**	**0.4110**	**0.8027**	**0.5191**	**0.4406**	0.7648

Table 2. Ablation study on different variants of our approach on the EndoVis-18 and EndoVis-17 datasets. Baseline, M1, M2, M3, M4, and M5 represent diverse ablation models, while Surgical-MambaLLM is our proposed model. Fusion Module represents the cross-modal fusion module in the network, Scanning Mode means different scanning modes utilized in the CBMI module

Models	Scanning Mode	Fusion Module	EndoVis-18			EndoVis-17		
			Acc	F-Score	mIoU	Acc	F-Score	mIoU
Baseline	×	×	0.6537	0.3595	0.7742	0.4216	0.3494	0.7315
M1	Simple 1D Scan	CBMI	0.6644	0.3335	0.7951	0.4826	0.3116	0.7434
M2	Bi-Scan [15]	CBMI	0.6615	0.3663	0.7915	0.4256	0.3774	0.7611
M3	Cross-Scan [19]	CBMI	0.6834	0.3420	0.7965	0.4675	0.3669	0.7348
M4	SIP	Mamba	0.6833	0.3795	0.7847	0.4778	0.4011	0.7506
M5	×	Transformer	0.6610	0.3524	0.7895	0.4766	0.3947	0.7559
Surgical-MambaLLM (our)	SIP	CBMI	**0.6964**	**0.4110**	**0.8027**	**0.5191**	**0.4406**	**0.7648**

Our LLM is based on InternLM-7B [27], featuring 32 layers. The projector is a 2-layer MLP. The classification head is a simple linear layer, while the location head is a 4-layer MLP. Our model is implemented using PyTorch and trained on a workstation equipped with 6 NVIDIA GeForce RTX 3090 GPUs. Following [4,3,5], we use Accuracy (Acc) [13], F-Score [13], and mIoU [25] as evaluation metrics. The batch size per GPU is 8. We utilize the AdamW optimizer [20] with a learning rate of 1×10^{-5} for the first stage and 1×10^{-6} for the second stage, and a dropout rate of 0.1.

3.3   Experiment Results

Comparisons with SOTAs. In Table 1, we compare the performance of Surgical-MambaLLM with other state-of-the-art models on the EndoVis-18 and EndoVis-17 datasets. The comprehensive results indicate that Surgical-MambaLLM outperforms other methods on both datasets, demonstrating its superiority. Specifically, Surgical-MambaLLM achieves significantly higher Accuracy, F-Score, and mIoU on the EndoVis-18 dataset compared to other methods. This indicates that Surgical-MambaLLM exhibits better visual understanding, visual reason-

ing, and localization capabilities when handling complex surgical scenes. On the EndoVis-17 dataset, Surgical-MambaLLM outperforms all SOTA methods in Accuracy and F-score, although its mIoU is a little lower than Surgical-VQLA++ [5]. This suggests that while Surgical-MambaLLM maintains strong scene understanding capabilities on the EndoVis-17 dataset, there is still room for improvement in localization prediction on the external validation dataset. Overall, through quantitative evaluation, we can conclude that our Surgical-MambaLLM possesses excellent visual reasoning and localization capabilities. Despite room for improvement in a certain metric, its overall performance significantly outperforms existing Surgical-VQLA methods, validating its superiority.

Ablation Studies. To validate the effectiveness of the components in our proposed Surgical-MambaLLM model, we conduct ablation study experiments on the EndoVis-18 and EndoVis-17 datasets, with the results presented in Table 2. First, we evaluate the effectiveness of the CBMI module by comparing the Baseline model with our Surgical-MambaLLM model. The baseline model directly concatenates textual and visual features as input to the LLM, bypassing the CBMI module. Compared to the baseline, Surgical-MambaLLM shows significant improvements across all metrics on both datasets, confirming the efficacy of the CBMI module.

Additionally, we explore the impact of different scanning modes on the model's performance by comparing M1, M2, and M3 with our Surgical-MambaLLM model. In M1, M2, and M3, the CBMI module employs different scanning methods: M1 employs simple 1D unidirectional scan, while M2 and M3 utilize the commonly used Bi-Scan mode [15] and Cross-Scan [19] mode. Our Surgical-MambaLLM employs the SIP scanning mode we designed. Compared to M1, M2, and M3, Surgical-MambaLLM achieves the best performance in terms of Accuracy, F-Score, and mIoU, validating the effectiveness of the SIP scanning mode.

Finally, we validate the effectiveness of Mamba2. In M4, Mamba2 is replaced with Mamba in CBMI while maintaining the same scanning mode; in M5, we employ the widely-used Cross-Attention [28] for feature fusion. Experimental results demonstrate that our model achieved superior performance in terms of accuracy, F-Score, and mIoU compared to both M4 and M5, thereby proving the superiority of Mamba2. In summary, these ablation studies demonstrate the critical role of each component in enhancing the overall performance of our Surgical-MambaLLM, underscoring the benefits of the CBMI module, SIP scanning mode, and Mamba2 model.

4   Conclusion

In this paper, we propose an innovative method, Surgical-MambaLLM, which is the first method integrating the Mamba model and LLM in Surgical-VQLA. To meet the unique characteristics of surgical scenarios, we designed the SIP scanning mode to comprehensively scan surgical images, enhancing the Mamba2 model's spatial awareness of surgical scenes. Additionally, we introduce the

CBMI module to achieve multimodal fusion, thereby improving the model's spatial understanding and cross-modal fusion capabilities for surgical images. Experiment results demonstrate that our Surgical-MambaLLM model outperforms SOTA methods on the EndoVis17-VQLA and EndoVis18-VQLA datasets, significantly improving the performance of the Surgical-VQLA task. Future work will focus on improving the model's performance on external validation sets and further improving the grounding capability. We plan to expand the dataset, optimize the model architecture, and refine training strategies to better leverage the potential of LLMs in the field of surgical robotics.

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

References

1. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
2. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
3. Bai, L., Islam, M., Ren, H.: Cat-vil: co-attention gated vision-language embedding for visual question localized-answering in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 397–407. Springer (2023)
4. Bai, L., Islam, M., Seenivasan, L., Ren, H.: 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 (2023)
5. Bai, L., Wang, G., Islam, M., Seenivasan, L., Wang, A., Ren, H.: Surgical-vqla++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery. Information Fusion **113**, 102602 (2025)
6. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2612–2620 (2017)
7. Ben-Younes, H., Cadene, R., Thome, N., Cord, M.: Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8102–8109 (2019)
8. Chen, K., Du, Y., You, T., Islam, M., Guo, Z., Jin, Y., Chen, G., Heng, P.A.: Llm-assisted multi-teacher continual learning for visual question answering in robotic surgery. arXiv preprint arXiv:2402.16664 (2024)
9. Dao, T., Gu, A.: Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. arXiv preprint arXiv:2405.21060 (2024)

10. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
11. Hao, P., Li, S., Wang, H., Kou, Z., Zhang, J., Yang, G., Zhu, L.: Surgery-r1: Advancing surgical-vqla with reasoning multimodal large language model via reinforcement learning. arXiv preprint arXiv:2506.19469 (2025)
12. Hao, P., Wang, H., Yang, G., Zhu, L.: Enhancing visual reasoning with llm-powered knowledge graphs for visual question localized-answering in robotic surgery. IEEE Journal of Biomedical and Health Informatics (2025)
13. Hossin, M., Sulaiman, M.N.: A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process **5**(2), 1 (2015)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
15. Jiang, X., Han, C., Mesgarani, N.: Dual-path mamba: Short and long-term bidirectional selective structured state space models for speech separation. arXiv preprint arXiv:2403.18257 (2024)
16. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960)
17. Krupa, A., Gangloff, J., Doignon, C., De Mathelin, M.F., Morel, G., Leroy, J., Soler, L., Marescaux, J.: Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing. IEEE transactions on robotics and automation **19**(5), 842–853 (2003)
18. Li, S., Ma, W., Guo, J., Xu, S., Li, B., Zhang, X.: Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12523–12533 (2024)
19. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model 2024. arXiv preprint arXiv:2401.10166 (2024)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article **2**(5) (2023)
22. Peters, B.S., Armijo, P.R., Krause, C., Choudhury, S.A., Oleynikov, D.: Review of emerging surgical robotic technology. Surgical endoscopy **32**, 1636–1655 (2018)
23. Qiao, Y., Yu, Z., Guo, L., Chen, S., Zhao, Z., Sun, M., Wu, Q., Liu, J.: Vlmamba: Exploring state space models for multimodal learning. arXiv preprint arXiv:2403.13600 (2024)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
25. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019)
26. Seenivasan, L., Islam, M., Krishna, A.K., Ren, H.: Surgical-vqa: Visual question answering in surgical scenes using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 33–43. Springer (2022)
27. Team, I.: Internlm: A multilingual language model with progressively enhanced capabilities (2023)

28. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
29. Wang, H., Jin, Y., Zhu, L.: Dynamic interactive relation capturing via scene graph learning for robotic surgical report generation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 2702–2709. IEEE (2023)
30. Wang, H., Yang, G., Zhang, S., Qin, J., Guo, Y., Xu, B., Jin, Y., Zhu, L.: Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. IEEE Transactions on Medical Imaging (2024)
31. Wu, H., Yang, Y., Xu, H., Wang, W., Zhou, J., Zhu, L.: Rainmamba: Enhanced locality learning with state space models for video deraining. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7881–7890 (2024)
32. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)
33. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems **29**(12), 5947–5959 (2018)
34. Zhao, H., Zhang, M., Zhao, W., Ding, P., Huang, S., Wang, D.: Cobra: Extending mamba to multi-modal large language model for efficient inference. arXiv preprint arXiv:2403.14520 (2024)
35. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)