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[3], Kaishun Wu[1], and Lei Zhu[1,2] (✉)

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

MICCAI 2025
28th International Conference on Medical Image Computing and Computer Assisted Intervention
23-27 September 2025
Daejeon Convention Center
Daejeon REPUBLIC OF KOREA

HKUST(GZ) HKUST

Paper Link:

Motivation

- Current methods primarily use Transformer-based approaches for cross-modal fusion, emphasizing global features and ***neglecting local details***. This makes it difficult to ***capture visual specifics and establish dependencies with the text***.
- LLMs still face significant challenges in understanding surgical scenes, particularly in ***perceiving spatial information*** due to the complexity of laparoscopic environments.

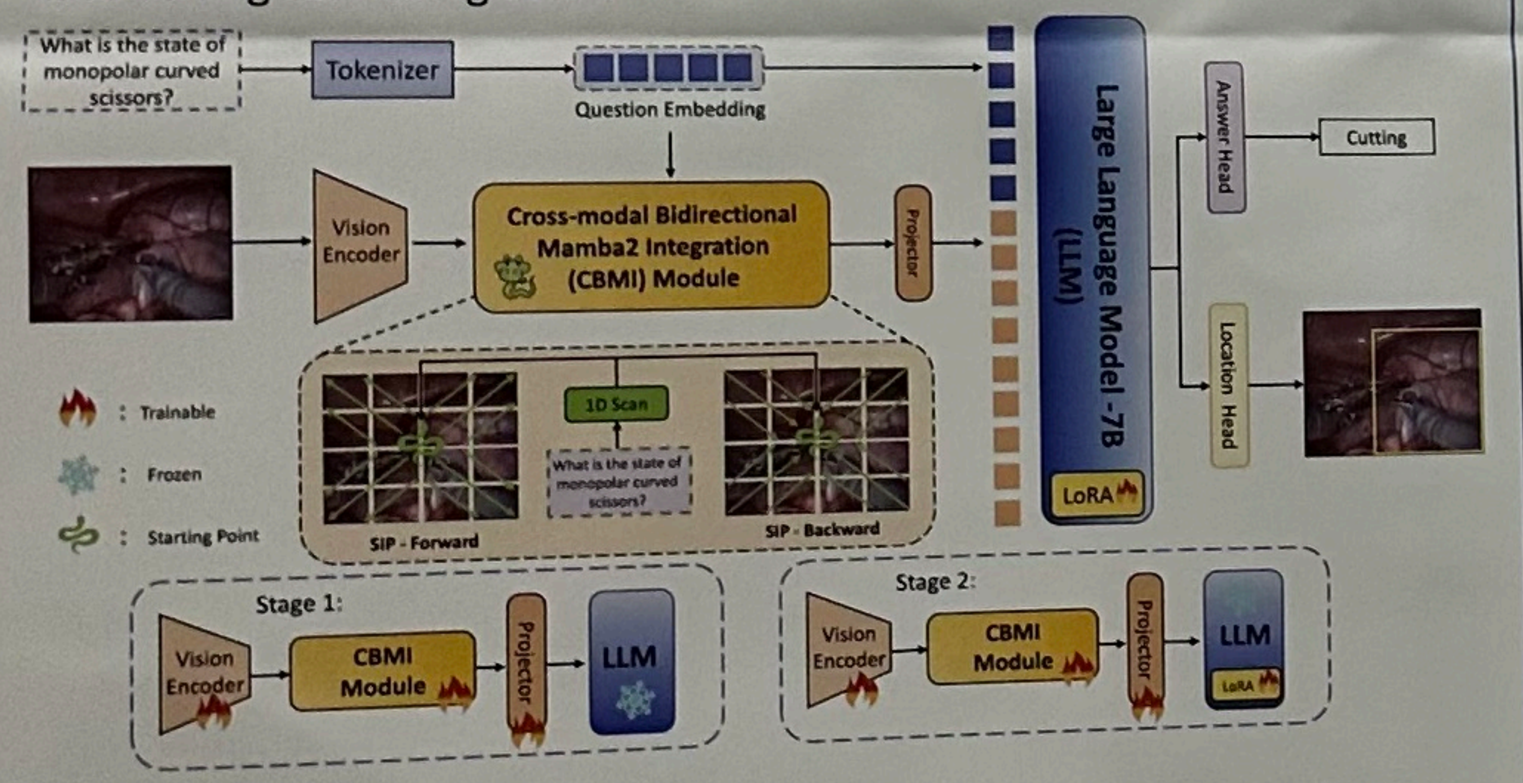
Contribution

- Surgical-MambaLLM is ***the first method to integrate Mamba2 with a Large Language Model for the surgical domain.***
- The CBMI module explores strategies to ***effectively merge visual and textual data within Mamba2.***
- The SIP mode improves Mamba2's ability to ***comprehend spatial aspects of surgical images.***
- Experiments reveal that Surgical-MambaLLM ***outperforms SOTA models.***

Method

• Overview of our Surgical-MambaLLM framework:

- 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.

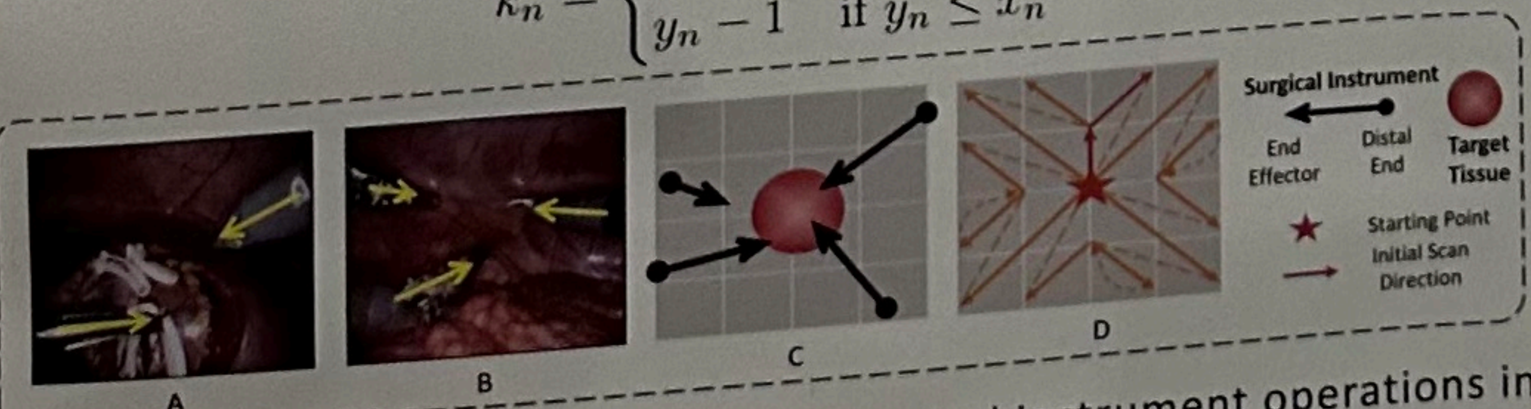


• Surgical Instrument Perception (SIP) Scanning Mode:

- 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. The trajectory can be described by the following formula:

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

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



- 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.

Method

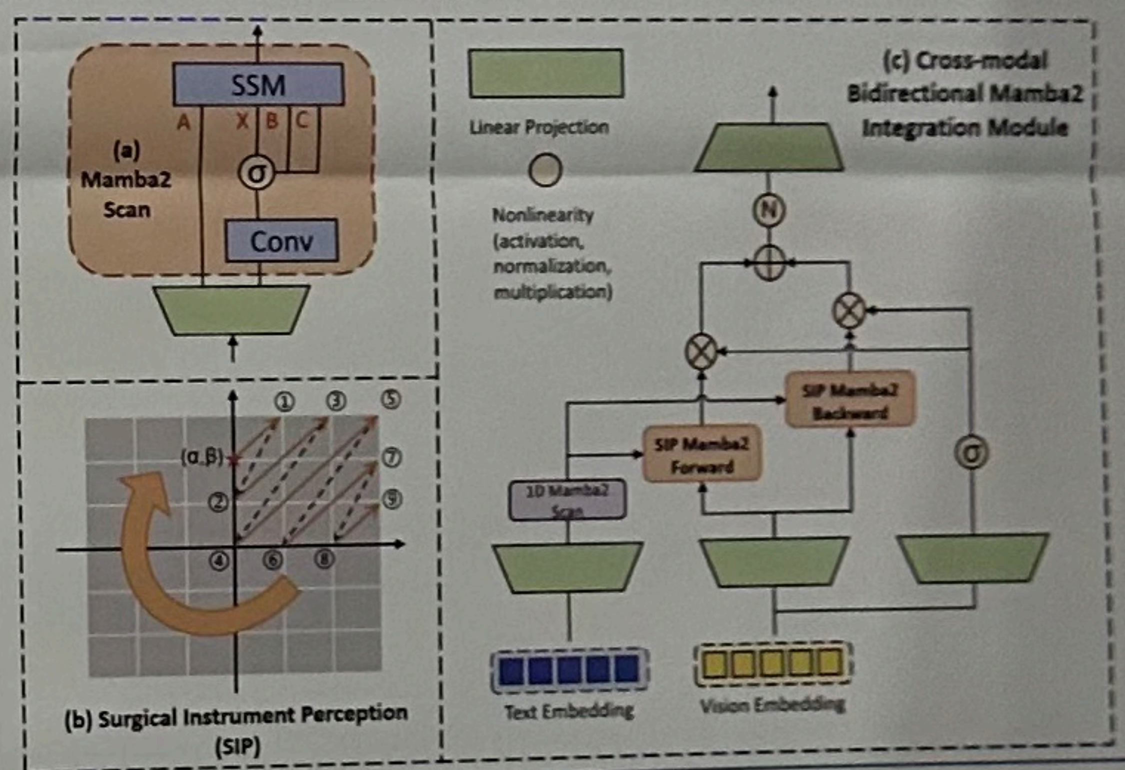
• Cross-modal Bidirectional Mamba2:

- 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. The formula is as follows:

$$F_t = l_t(t), F_v = l_v(v),$$

$$S_{\text{forward}} = \text{SIP-Mamba2}_{\text{forward}}(F_t, F_v), \quad S = S_{\text{forward}} \cdot \sigma(F_v) + S_{\text{backward}} \cdot \sigma(F_v),$$

$$S_{\text{backward}} = \text{SIP-Mamba2}_{\text{backward}}(F_t, F_v), \quad S_{\text{output}} = Linear(LN(S)).$$



Experimental Results

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

Ablation study on different variants of our approach on the EndoVis-18 and EndoVis-17 datasets.

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**