Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Question guided progressive reasoning for surgical visual question localized answering

Lianliang Li
Beijing Institute of Technology
Xue Li
Beijing Institute of Technology
Weiyan Kong
Beijing Institute of Technology
Zhaoyun Wu
Beijing Institute of Technology
Changsheng Li
Beijing Institute of Technology
Xingguang Duan
duanstar@bit.edu.cn

Beijing Institute of Technology

Article

Keywords: Robotic-assisted surgery, Mamba, Surgical visual question localized-answering, Multi-hop reasoning, vision-language embedding

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11350613/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Question guided progressive reasoning for surgical visual question localized answering

Lianliang Li[1], Xue Li[1], Weiyan Kong[1], Zhaoyun Wu[1], Changsheng Li[1,2,3*], Xingguang Duan[1,2,3*]

[1]School of Mechatronical Engineering, Beijing Institute of Technology, No. 5, South Zhongguancun Street, Haidian District, Beijing, 100081, China.
[2]Key Laboratory of Biomimetic Robots and Systems, Ministry of Education, Beijing Institute of Technology, No. 5, South Zhongguancun Street, Haidian District, Beijing, 100081, China.
[3]School of Medical Engineering, Beijing Institute of Technology, No. 6 Jinfeng Road, Zhuhai, 519088, China.

*Corresponding author(s). E-mail(s): lics@bit.edu.cn; duanstar@bit.edu.cn;

Abstract

Surgical visual question localized answering (VQLA) connects language-guided scene understanding with spatial perception for context-aware surgical robotics. However, existing methods often use fixed multimodal representations, limiting adaptation of visual evidence to each question's semantic and spatial needs. We propose SurgPR-Mamba, a question-guided progressive reasoning framework that combines Mamba-based multimodal modeling with progressive evidence refinement. Spatially aware visual features are aligned with question semantics and integrated through cross-modal Mamba interaction to establish question-relevant representations while preserving contextual dependencies. Based on this representation, visual evidence and question states are iteratively updated across multiple reasoning hops, allowing the reasoning focus to evolve from broader surgical context towards more discriminative target- and interaction-related cues. The refined evidence is subsequently used for joint answer prediction and visual localization. With 174.03M parameters, SurgPR-Mamba achieves competitive performance on EndoVis-18-VQLA and reliable generalization on EndoVis-17-VQLA. These results suggest that question-guided progressive reasoning may provide a compact and spatially grounded approach to surgical scene understanding for context-aware robotic assistance.

Keywords: Robotic-assisted surgery, Mamba, Surgical visual question localized-answering, Multi-hop reasoning, vision-language embedding

Introduction

Robot-assisted minimally invasive surgery requires robust scene understanding to support increasingly capable assistance and autonomy [1–4]. Surgical visual question answering (VQA) connects visual perception with language, enabling models to answer questions about anatomical structures, instrument identity and location, and surgical states [5–7]. Surgical visual question localized answering (VQLA) extends this task by pairing each answer with a localized image region, linking semantic predictions to their visual evidence [8, 9]. Establishing this correspondence is challenging in surgical scenes, where visually similar instruments and anatomical structures, frequent occlusions, and complex instrument–tissue interactions can obscure the evidence needed to answer a question. Moreover, that evidence may occupy only a small region of the operative field, making global scene recognition insufficient [10]. Surgical VQLA therefore requires models to identify question-relevant visual cues and reason about them in context to jointly support accurate answering and localization.

Advances in surgical VQA and VQLA have improved multimodal scene understanding through gated vision–language fusion, fine-grained cross-modal alignment, and joint answer prediction and localization. Subsequent research efforts have concentrated on improving robustness, integrating answer semantics, and incorporating external knowledge to bolster visual reasoning and grounding capabilities [11–13]. These innovations have predominantly aimed at developing more comprehensive multimodal representations, yet there has been comparatively limited focus on the explicit maintenance and iterative updating of question-relevant visual evidence throughout sequential reasoning stages. Selective state-space models [14, 15], exemplified by Mamba, provide a framework for propagating contextual information via input-dependent state transitions [16]. Their applications to visual and surgical vision–language tasks suggest potential for modeling spatial context and cross-modal dependencies [17]. However, selective state-space models in surgical VQLA have primarily been used for feature representation and multimodal fusion, leaving their potential for question-guided evidence refinement less explored. This motivates a progressive reasoning framework that updates visual evidence under question guidance and uses the evidence acquired at each step to inform subsequent reasoning.

Answering surgical questions requires identifying relevant visual evidence and interpreting it in the context of the question. During surgery, clinicians often interpret the operative field by progressively directing attention from the broader scene to a specific instrument and its interaction with tissue [18–20]. This progression from contextual assessment to focused interpretation motivates the question-guided reasoning strategy illustrated in Fig. 1. Take the question "What is the state of the ProGrasp forceps?" as an example. To provide a reliable answer, the model must first understand the overall surgical scene, then locate the specific ProGrasp forceps among possibly several instruments, and finally analyze how it interacts with nearby tissue to determine its functional state [21, 22]. The visual evidence required changes across reasoning stages, from contextual cues that establish the operative scene to instrument-specific features that identify the target and local interaction cues that inform its functional state [23, 24]. A single fused representation may not adequately distinguish these complementary roles or preserve their relevance to the question. This motivates question-guided progressive reasoning, in which visual evidence is successively selected, refined and integrated to support the final answer.

Implementing progressive reasoning within surgical VQLA necessitates more than merely achieving precise alignment between images and corresponding questions. Surgical scenes are complex and contain a lot of visual information that may not be relevant to the specific question, so it is crucial to develop multimodal representations that are conditioned on the question, preserving spatial details while focusing on evidence pertinent to the query. Although OTAS improves fine-grained alignment and incorporates answer semantics into prediction [25], these mechanisms do not explicitly define how evidence acquired at one reasoning step should guide visual processing at the next. The challenge is therefore to preserve spatially precise evidence while allowing its interpretation to evolve through interaction with the question. SSG-VQA enriches scene understanding with object-level features and spatial and action relationships [21], yet representing these relationships leaves a further challenge of determining which cues should be revisited as evidence accumulates. Surgery-R1 addresses reasoning through chain-of-thought supervision and reinforcement learning with a multimodal coherence reward [26], but evaluating the resulting reasoning and predictions does not by itself establish how intermediate visual evidence should be maintained and updated. The remaining challenge is to couple evidence selection with an evolving question state while preserving spatial detail and the original intent of the query. This coupling must accommodate the distinct evidence requirements of questions about anatomy, instrument identity, location and state, and support both answer discrimination and localization. These requirements motivate a unified framework for question-conditioned representation learning, progressive evidence refinement and question-aware prediction.

Clinically relevant information is distributed across different semantic levels, requiring informative visual cues and their relationships to be captured as reasoning progresses. Surgical questions may focus on anatomy, instrument identity, spatial location, or action state, with each type relying on different combinations of visual evidence and semantic context [27–29]. Intermediate evidence can therefore be progressively refined to support subsequent reasoning while maintaining consistency with the original question intent [30]. Such representations further facilitate alignment between the predicted answer and the corresponding localized evidence. These considerations motivate question-guided multimodal representation learning, progressive evidence refinement, and question-aware prediction.

To address the above challenges, SurgPR-Mamba, a question-guided progressive reasoning framework, is proposed for surgical VQLA. Rather than directly aggregating visual and linguistic information into a fixed multimodal representation, SurgPR progressively identifies, updates, and exploits question-relevant visual evidence throughout the reasoning process. By integrating selective state-space modeling into multimodal representation and evidence reasoning, SurgPR enables the required visual evidence to evolve from coarse contextual cues toward more discriminative target- and interaction-level information.

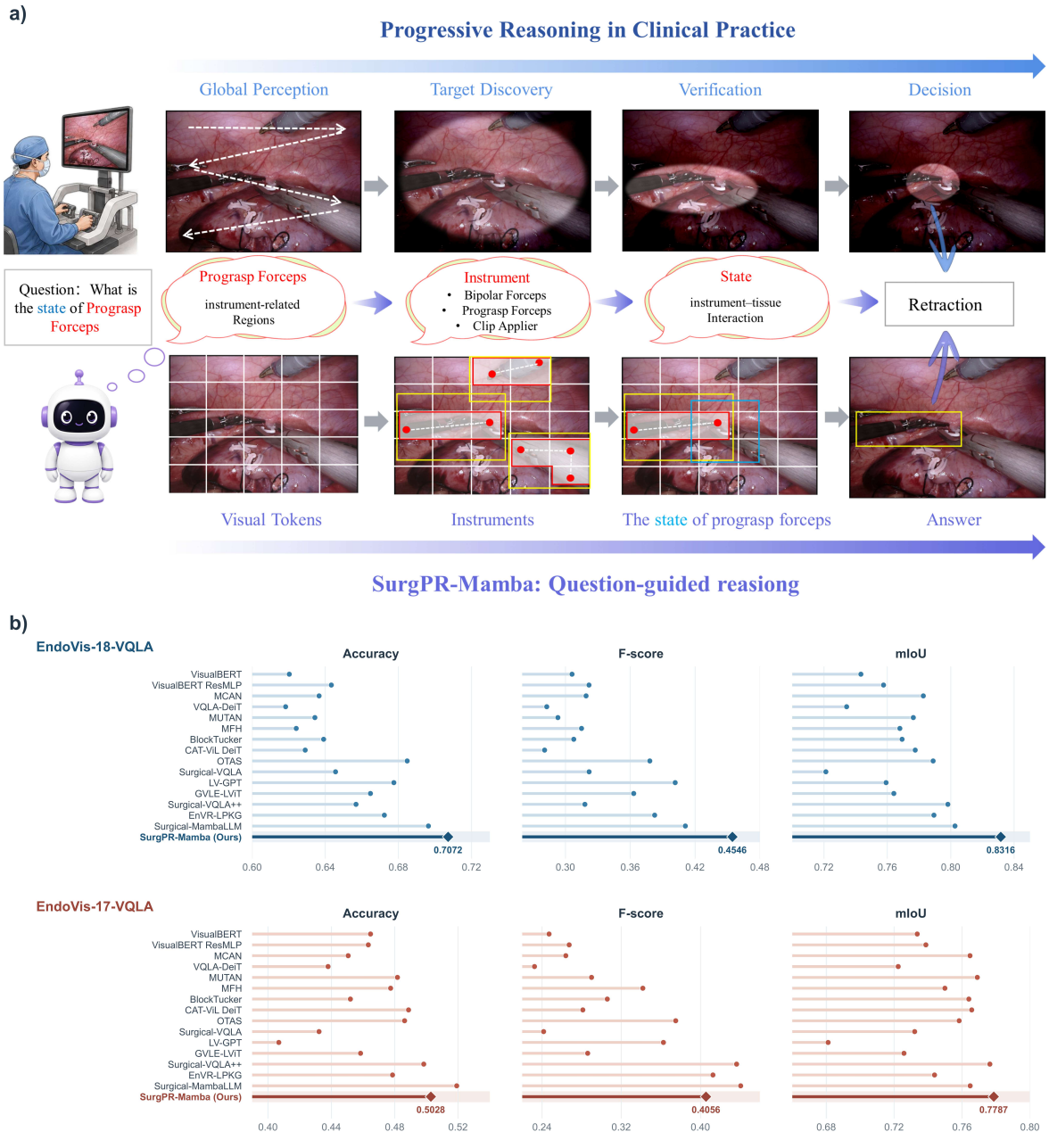


Fig. 1 Progressive reasoning in surgical visual question localized answering and benchmark performance. a) Original conceptual illustration of progressive reasoning: clinical perception proceeds from global context through target discovery and verification to a decision, while SurgPR-Mamba uses question-guided visual evidence to predict an answer and its location. The illustration motivates the design rather than establishing equivalence between model hops and clinical reasoning stages. b) Accuracy, F-score, and mIoU of all 16 methods in Table 1. The upper (blue) and lower (terracotta) rows show EndoVis-18-VQLA and EndoVis-17-VQLA, respectively; columns correspond to the three metrics. Dark diamonds and annotated scores highlight SurgPR-Mamba. Methods follow table order. Horizontal ranges are adjusted separately for each panel to reveal differences; marker positions encode scores, and higher scores indicate better performance.

The main contributions of this work are summarized as follows:

1. A Question-guided Mamba Representation (QMR) module has been created to improve the extraction of question-relevant visual information and to build effective cross-modal representations for complex surgical scenes.

2. A Progressive Evidence Reasoning (PER) approach is proposed to iteratively refine visual evidence through multiple reasoning steps, allowing the model to move from a global context to evidence at the target and interaction levels for precise answer localization.

3. Extensive experiments demonstrate the effectiveness of SurgPR-Mamba on EndoVis-18-VQLA and its competitive cross-dataset performance on EndoVis-17-VQLA.

Table 1 | Comparison with state-of-the-art methods on the EndoVis-18-VQLA and EndoVis-17-VQLA datasets. The best results are shown in bold, the second-best results are underlined.

Method	EndoVis-18-VQLA			EndoVis-17-VQLA		
	Acc	F-Score	mIoU	Acc	F-Score	mIoU
VisualBERT [6]	0.6203	0.3061	0.7433	0.4647	0.2471	0.7336
VisualBERT ResMLP [6]	0.6433	0.3218	0.7576	0.4633	0.2674	0.7387
MCAN [31]	0.6366	0.3191	0.7827	0.4506	0.2640	0.7647
VQLA-DeiT [32]	0.6183	0.2827	0.7343	0.4379	0.2324	0.7224
MUTAN [33]	0.6343	0.2930	0.7764	0.4817	0.2900	0.7690
MFH [34]	0.6241	0.3149	0.7679	0.4774	0.3418	0.7499
BlockTucker [35]	0.6391	0.3077	0.7693	0.4520	0.3059	0.7640
CAT-ViL DeiT [11]	0.6290	0.2807	0.7776	0.4887	0.2812	0.7657
OTAS [25]	0.6848	0.3781	0.7889	0.4862	0.3750	0.7583
Surgical-VQLA [8]	0.6456	0.3219	0.7212	0.4322	0.2416	0.7321
LV-GPT [36]	0.6775	0.4016	0.7592	0.4068	0.3626	0.6811
GVLE-LViT [8]	0.6647	0.3632	0.7641	0.4584	0.2861	0.7258
Surgical-VQLA++ [9]	0.6568	0.3180	0.7982	0.4983	<u>0.4365</u>	<u>0.7764</u>
EnVR-LPKG [10]	0.6723	0.3826	0.7894	0.4786	0.4126	0.7438
Surgical-MambaLLM [17]	<u>0.6964</u>	<u>0.4110</u>	<u>0.8027</u>	**0.5191**	**0.4406**	0.7648
SurgPR-Mamba (Ours)	**0.7072**	**0.4546**	**0.8316**	<u>0.5028</u>	0.4056	**0.7787**

Results

Benchmark comparison

On EndoVis-18-VQLA, SurgPR-Mamba achieves the best overall performance across all three evaluation metrics, as shown in Fig. 1b and Table 1. Compared with Surgical-VQLA++, Acc, F-Score, and mIoU are improved by 5.04, 13.66, and 3.34 percentage points, respectively. The larger gain in F-Score is particularly relevant to surgical VQLA, where many answer categories are visually similar and unevenly distributed. In such cases, recognizing the general surgical content is often insufficient, and the model must distinguish subtle differences in instrument state, location, or interaction. SurgPR-Mamba addresses this by using question semantics to restrict the relevant visual evidence and by refining this evidence over successive reasoning steps. This also benefits localization, as the predicted region is more closely tied to the information required for answering the question. Compared with Surgical-MambaLLM, SurgPR-Mamba further improves Acc, F-Score, and mIoU by 1.08, 4.36, and 2.89 percentage points, respectively. The results across these comparisons highlight the importance of jointly improving semantic discrimination and visual grounding, rather than relying only on global multimodal representations for answer prediction.

On EndoVis-17-VQLA, which is excluded from training, SurgPR-Mamba retains localization performance when transferred to unseen surgical scenes. The mIoU reaches 0.7787, compared with 0.7764 for Surgical-VQLA++ and 0.7648 for Surgical-MambaLLM. However, Surgical-MambaLLM has higher accuracy (0.5191 versus 0.5028) and F-Score (0.4406 versus 0.4056), so the transfer advantage does not extend to every answer-prediction metric. In contrast, answer prediction remains more challenging under the cross-dataset shift, particularly for F-Score. This difference indicates that spatial evidence transfers more consistently than category-level semantics across datasets. By aligning the question with local visual regions and repeatedly refining the selected evidence, SurgPR-Mamba reduces its dependence on scene-specific visual patterns. The results therefore support the value of question-guided evidence reasoning for preserving reliable visual grounding under changes in surgical appearance and data distribution.

The qualitative comparisons in Fig. 2 further illustrate how different models use visual evidence to support their predictions. The differences become more apparent for questions involving instrument state, location, and identity, where recognizing the presence of an instrument alone is insufficient. For example, several methods can identify the queried instrument but confuse its action state, such as *Idle*, *Retraction*, or *Cutting*. Similar errors occur in location and instrument-identification questions, where the predicted region is related to the surgical scene but does not fully correspond to the target specified by the question. In contrast, SurgPR-Mamba shows more consistent agreement between the predicted answer and the localized region across these cases. Its localization also tends to remain concentrated around the evidence that is directly relevant to the question, rather than surrounding visually salient

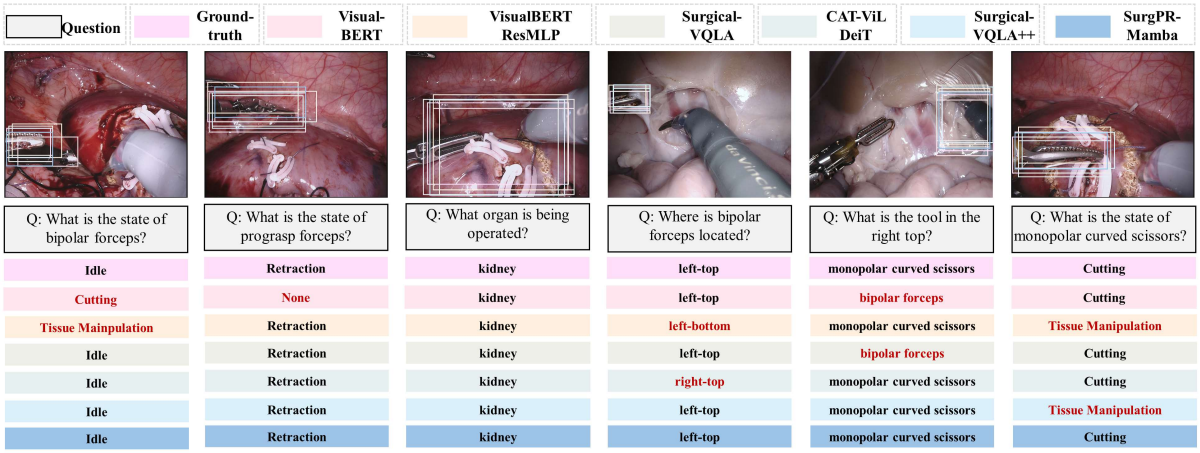


Fig. 2 Qualitative comparison of surgical VQLA results across different question types. The colored bounding boxes show the localization results of different methods, while the corresponding answer predictions are presented below each image. Red text indicates incorrect predictions.

but less informative regions. This difference is important in surgical VQLA because similar instruments may appear simultaneously and their states often depend on subtle local appearance or interactions with surrounding tissue [37]. The progressive reasoning process allows the model to repeatedly reassess the relevant visual information as the question representation evolves. As a result, SurgPR-Mamba can move from coarse target identification toward more specific evidence related to state, position, or interaction, providing a clearer connection between visual grounding and the final answer.

Question-type Analysis. To examine whether the overall gains are consistent across distinct semantic requirements, question-type performance is compared among CAT-ViL DeiT, Surgical-VQLA, Surgical-VQLA++, and SurgPR-Mamba in Table 2. The advantages of SurgPR-Mamba are most evident for questions that require fine-grained evidence discrimination. On EndoVis-18-VQLA, state questions achieve 0.6871 Acc and 0.3499 F-Score, exceeding Surgical-VQLA++ by 11.00 and 14.77 percentage points, respectively, while mIoU further increases to 0.8944. Location questions show a similar pattern, with F-Score improving from the previous best of 0.5616 to 0.6567. These gains suggest that progressively refining question-relevant evidence is particularly beneficial when prediction depends on spatial relations, local appearance, or instrument tissue interactions rather than object presence alone. For organ questions, where answer recognition is already saturated, the improvement is mainly reflected in localization, with mIoU increasing from 0.8124 to 0.8548, indicating more precise spatial grounding. Instrument questions show smaller gains on EndoVis-18-VQLA but transfer more clearly to EndoVis-17-VQLA, where SurgPR-Mamba reaches 0.6212 Acc, 0.4843 F-Score, and 0.8456 mIoU. This suggests that question-guided evidence refinement can preserve discriminative object-level cues under changes in surgical scenes. Although state and location classification remain more sensitive to cross-dataset variation, the overall results indicate that the proposed reasoning strategy is better suited to the heterogeneous evidence requirements of different surgical questions than a single fixed multimodal representation.

Efficiency Analysis. The model efficiency is further evaluated by comparing the number of trainable parameters with representative surgical VQLA methods, as shown in Fig. 3. SurgPR-Mamba contains 174.03M parameters, substantially fewer than CAT-ViL DeiT (897.60M), Surgical-VQLA (826.45M), and Surgical-VQLA++ (897.60M). Compared with Surgical-VQLA++, the parameter count is reduced by approximately 80.6%, corresponding to a 5.16× smaller model. The parameter comparison does not establish inference latency, peak memory use or clinical deployment feasibility; these require matched measurements. This reduction mainly benefits from the selective state-space formulation, which models contextual dependencies without relying on large-scale attention-based interaction throughout the multimodal pipeline. Despite the considerably smaller model size, SurgPR-Mamba achieves higher F-Score and mIoU on EndoVis-18-VQLA and maintains competitive cross-dataset performance. These results indicate that progressive evidence reasoning can improve the utilization of multimodal information without requiring a substantial increase in model complexity, providing a more efficient balance between representation capacity and surgical VQLA performance.

Ablation studies

An overall ablation study is conducted to examine the contribution of the main components in SurgPR-Mamba. As shown in Table 3, the baseline retains the basic visual and linguistic representations, token

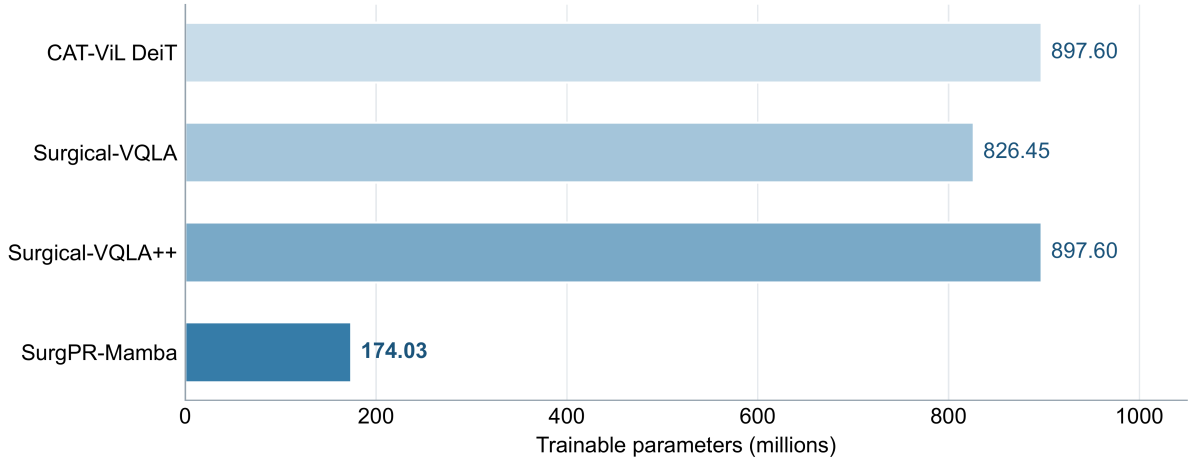


Fig. 3 Comparison of model parameter counts among representative surgical VQLA methods.

Table 2| Comparison of different methods across question types on EndoVis-18-VQLA and EndoVis-17-VQLA. The best results are highlighted in bold.

Method	Data Type	EndoVis-18-VQLA			EndoVis-17-VQLA		
		Acc	F-Score	mIoU	Acc	F-Score	mIoU
CAT-ViL DeiT	Organ	1	1	0.6910	–	–	–
Surgical-VQLA		1	1	0.7253	–	–	–
Surgical-VQLA++		1	1	0.8124	–	–	–
SurgPR-Mamba		1	1	**0.8548**	–	–	–
CAT-ViL DeiT	State	0.5426	0.1646	0.8279	0.3305	0.1898	0.7656
Surgical-VQLA		0.5049	0.1676	0.8174	0.3051	0.2882	0.7479
Surgical-VQLA++		0.5771	0.2022	0.8655	0.2966	0.3336	**0.7972**
SurgPR-Mamba		**0.6871**	**0.3499**	**0.8944**	**0.3432**	**0.3523**	0.7505
CAT-ViL DeiT	Location	0.5585	0.4802	0.7491	0.5297	0.4988	0.7182
Surgical-VQLA		0.6117	0.5616	0.6411	0.5424	0.5269	0.6450
Surgical-VQLA++		0.6064	0.4678	0.7515	**0.6441**	**0.6181**	0.7187
SurgPR-Mamba		**0.6631**	**0.6567**	**0.7691**	0.5424	0.5142	**0.7301**
CAT-ViL DeiT	Instrument	0.6791	0.2266	0.7911	0.5763	0.4664	0.8176
Surgical-VQLA		0.6722	0.2067	0.7039	0.5593	0.4002	0.7166
Surgical-VQLA++		0.6543	0.2194	0.7820	0.5805	0.4173	0.8116
SurgPR-Mamba		**0.6845**	**0.2321**	**0.8138**	**0.6212**	**0.4843**	**0.8456**

interaction, and prediction heads, while removing the proposed question-guided representation and progressive reasoning components. Introducing QMR increases Acc, F-Score, and mIoU on EndoVis-18-VQLA by 5.33, 5.49, and 6.10 percentage points, respectively. A similar improvement is observed on EndoVis-17-VQLA, particularly for localization. QMR preserves the spatial organization of visual evidence and explicitly associates question semantics with relevant surgical regions. This provides a stronger basis for both semantic prediction and spatial grounding than global multimodal interaction alone.

The inclusion of PER further strengthens semantic reasoning on EndoVis-18-VQLA, increasing Acc and F-Score to 0.6807 and 0.4304, respectively, while preserving a comparable level of spatial grounding. On EndoVis-17-VQLA, adding PER at this intermediate stage decreases all three metrics relative to QMR alone, so improvements are not monotonic across configurations. This indicates that progressive reasoning mainly benefits the interpretation of question-relevant evidence by incorporating contextual and relational cues across successive reasoning steps, rather than simply concentrating the predicted region. After QMP is introduced, the complete model further improves to 0.7072 Acc, 0.4546 F-Score, and 0.8316 mIoU, showing that the refined evidence can be more effectively translated into coordinated semantic and localization predictions. A similar effect is observed under cross-dataset evaluation, where QMP raises

Table 3| Overall ablation study of the proposed components on EndoVis-18-VQLA and EndoVis-17-VQLA. The best results are highlighted in bold.

Model Configuration				EndoVis-18-VQLA			EndoVis-17-VQLA		
Baseline	QMR	PER	QMP	Acc	F-Score	mIoU	Acc	F-Score	mIoU
✓				0.6260	0.3619	0.7695	0.4788	0.3600	0.7269
✓	✓			0.6793	0.4168	0.8305	0.4844	0.3771	0.7768
✓	✓	✓		0.6807	0.4304	0.8291	0.4802	0.3661	0.7747
✓	✓	✓	✓	**0.7072**	**0.4546**	**0.8316**	**0.5028**	**0.4056**	**0.7787**

Table 4| Ablation study of the components in QMR on EndoVis-18-VQLA and EndoVis-17-VQLA. SME denotes Spatial-aware Mamba Encoding, QEA denotes Question-guided Evidence Alignment, and CMF denotes Cross-modal Mamba Fusion. The best results are highlighted in bold.

QMR Modules			EndoVis-18-VQLA			EndoVis-17-VQLA		
SME	QEA	CMF	Acc	F-Score	mIoU	Acc	F-Score	mIoU
✓			0.6575	0.3946	0.8257	0.4689	0.3456	0.7701
✓	✓		0.6615	0.4145	0.8287	0.4661	0.3300	0.7743
✓	✓	✓	**0.7072**	**0.4546**	**0.8316**	**0.5028**	**0.4056**	**0.7787**

the performance on EndoVis-17-VQLA to 0.5028 Acc, 0.4056 F-Score, and 0.7787 mIoU. These results suggest that QMR, PER, and QMP form a complementary reasoning pipeline in which question-relevant evidence is first established, progressively refined, and finally converted into more consistent answer and localization outputs.

Question-guided Mamba Representation analysis

To evaluate the role of QMR in question evidence alignment, its three components are examined in Table 4. For the internal ablation of QMR, PER and QMP are kept unchanged, while only the individual components within QMR are varied. Spatial-aware Mamba Encoding (SME) retains spatial information during visual encoding and provides the initial representation for subsequent multimodal interaction. After Question-guided Evidence Alignment (QEA) is introduced, the F-Score on EndoVis-18-VQLA increases from 0.3946 to 0.4145, while Acc and mIoU also show modest improvements from 0.6575 to 0.6615 and from 0.8257 to 0.8287, respectively. This suggests that QEA primarily enhances the relevance of visual evidence to the question, improving answer discrimination without substantially altering the underlying spatial representation. On EndoVis-17-VQLA, although Acc and F-Score exhibit slight fluctuations, mIoU increases from 0.7701 to 0.7743, indicating that question conditioning can still contribute to spatial evidence refinement, while its effect on answer prediction is less stable under cross-dataset evaluation.

The introduction of Cross-modal Mamba Fusion (CMF) further improves both answer prediction and localization. On EndoVis-18-VQLA, Acc, F-Score, and mIoU reach 0.7072, 0.4546, and 0.8316, respectively. Similar improvements are observed on EndoVis-17-VQLA, where Acc and F-Score increase to 0.5028 and 0.4056, while mIoU further improves to 0.7787. These results highlight the role of Mamba in modeling contextual dependencies beyond local question visual correspondence. Through selective state-space modeling, CMF propagates question-related information across visual tokens and integrates complementary linguistic context into the visual stream, allowing locally aligned evidence to be interpreted within the broader surgical scene. Meanwhile, the parallel visual and textual branches preserve modality-specific information, while cross-modal interaction regulates the information exchanged between them. This combination is particularly beneficial for answer prediction while maintaining reliable spatial localization, as identifying the queried target often depends not only on local appearance but also on its semantic relation to the question and surrounding context. Together, spatial encoding, question-guided alignment, and Mamba-based cross-modal interaction provide a more coherent question-aware representation for subsequent progressive reasoning.

Progressive Evidence Reasoning analysis

Effect of Visual Update. To determine whether progressive reasoning benefits from dynamically refining the visual evidence, the visual and question updates in PER are examined separately in Table 5. For the

Table 5| Ablation study of the visual and question update mechanisms in PER on EndoVis-18-VQLA and EndoVis-17-VQLA. The best results are highlighted in bold.

PER Modules		EndoVis-18-VQLA			EndoVis-17-VQLA		
Visual Update	Question Update	Acc	F-Score	mIoU	Acc	F-Score	mIoU
×	×	0.6780	0.4346	0.8278	0.4659	0.3537	0.7704
✓	×	0.6876	0.4304	0.8291	0.4802	0.3661	0.7747
✓	✓	**0.7072**	**0.4546**	**0.8316**	**0.5028**	**0.4056**	**0.7787**

Table 6| Effect of different reasoning hops on EndoVis-18-VQLA and EndoVis-17-VQLA. The best results are highlighted in bold.

Hops	EndoVis-18-VQLA			EndoVis-17-VQLA		
	Acc	F-Score	mIoU	Acc	F-Score	mIoU
1	0.6599	0.3832	0.8322	0.4985	0.3852	0.7770
2	0.6756	0.3906	**0.8344**	0.4731	0.3517	0.7729
3	**0.7072**	**0.4546**	0.8316	**0.5028**	**0.4056**	**0.7787**
4	0.6628	0.3849	0.8262	0.4929	0.3614	0.7682

internal ablation of PER, QMR and QMP are kept fixed, while the visual and question update mechanisms are varied. Introducing visual update increases Acc from 0.6780 to 0.6876 and mIoU from 0.8278 to 0.8291 on EndoVis-18-VQLA, although the F-Score shows a slight decrease from 0.4346 to 0.4304. On EndoVis-17-VQLA, all three metrics improve, with Acc, F-Score, and mIoU increasing from 0.4659, 0.3537, and 0.7704 to 0.4802, 0.3661, and 0.7747, respectively. These results indicate that the evidence required by a question is not fully captured by the initial multimodal representation. Re-evaluating the visual features according to the current question state helps the model retain more relevant spatial cues and provides more reliable evidence for subsequent reasoning. The improvement in localization and cross-dataset performance suggests that dynamically refined visual evidence is more effective than maintaining a fixed representation throughout the reasoning process.

Effect of Question Update. The question update is further introduced to examine whether the refined visual evidence can benefit subsequent reasoning when it is fed back into the question state. Combining the two updates further improves all three metrics on both datasets. On EndoVis-18-VQLA, Acc, F-Score, and mIoU increase to 0.7072, 0.4546, and 0.8316, respectively, while the corresponding results on EndoVis-17-VQLA reach 0.5028, 0.4056, and 0.7787. The more pronounced gains in Acc and F-Score indicate that feeding the refined visual evidence back into the question representation strengthens semantic discrimination in subsequent reasoning steps. In this process, the question state not only guides visual evidence selection but is also progressively revised by the evidence obtained at each hop. This reciprocal interaction allows later hops to reconsider the surgical scene with an updated semantic state, supporting the formulation of PER as an iterative refinement between visual evidence and question semantics rather than a sequence of independent visual updates.

Reasoning analysis

Effect of Multi-hop Reasoning. To verify whether multi-hop reasoning can progressively improve the use of question-related visual evidence, different reasoning depths are compared in Fig. 5(a) and Table 6. On EndoVis-18-VQLA, increasing the reasoning depth from one to three hops improves Acc from 0.6599 to 0.7072 and F-Score from 0.3832 to 0.4546. Although mIoU reaches its highest value of 0.8344 at two hops and slightly decreases to 0.8316 at three hops, the three-hop configuration provides substantially better answer prediction while maintaining comparable localization performance. A similar overall pattern is observed on EndoVis-17-VQLA. Despite some fluctuations at two hops, three-hop reasoning achieves the highest Acc, F-Score, and mIoU of 0.5028, 0.4056, and 0.7787, respectively. These results suggest that a single reasoning step may not fully capture the evidence required by the question, whereas successive interactions allow the model to progressively reconsider and refine question-relevant information. When the reasoning depth is further increased to four hops, both answer prediction and localization deteriorate on the two datasets, indicating that excessive refinement may introduce redundant or noisy information

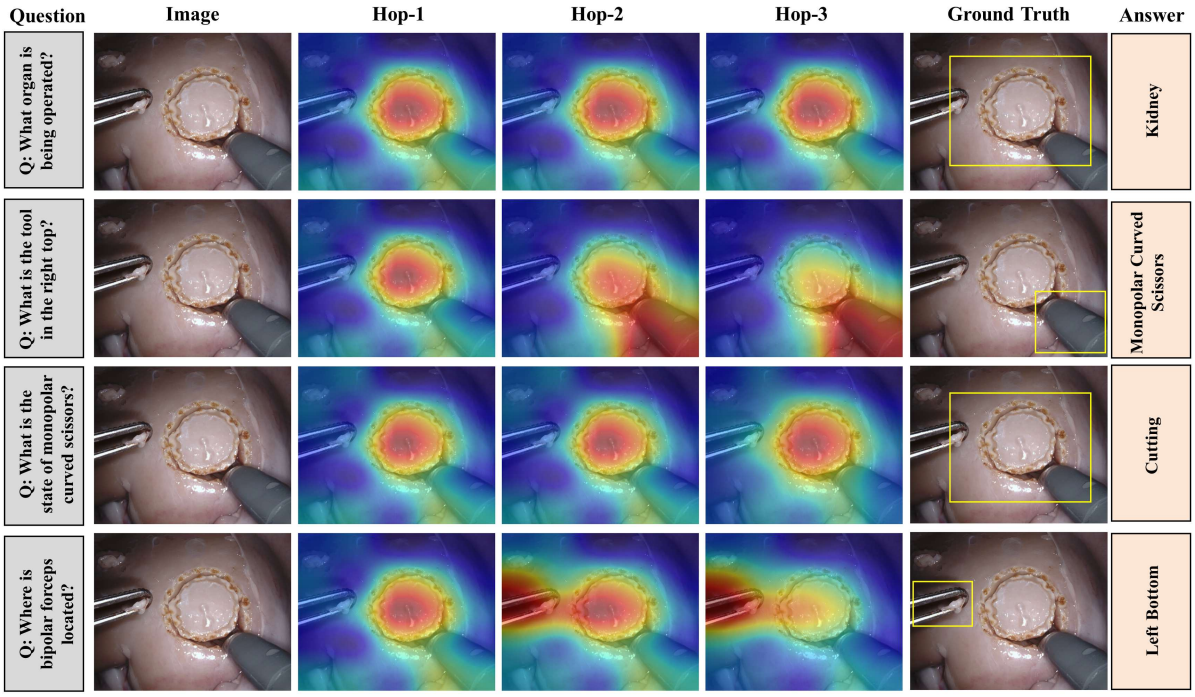


Fig. 4 Visualization of question-conditioned evidence evolution across different question types in the same surgical scene. Evidence maps from three reasoning hops are shown for organ, instrument, state, and location questions, together with the corresponding ground-truth regions and predicted answers.

and disturb evidence accumulated in earlier hops. Overall, three-hop reasoning provides the most favorable balance between semantic prediction and visual grounding.

Evolution of Visual Evidence. To further understand how visual evidence evolves during multi-hop reasoning, Fig. 6 visualizes the evidence maps for representative questions from different semantic categories. Despite differences in surgical scenes and queried targets, a common tendency can be observed across the reasoning process. Early responses are relatively distributed and often cover several potentially relevant structures, whereas subsequent hops gradually reorganize the evidence around regions that are more closely related to the question. The resulting patterns also differ with the information required for answering. Organ-related questions retain relatively broad anatomical context, while questions concerning instrument identity or location show more concentrated responses around the queried instrument. For state-related questions, the informative evidence may further involve local regions surrounding the instrument, where its interaction with the surgical field can provide additional cues. The progression therefore does not simply correspond to increasingly compact attention. Instead, the visual evidence is gradually reorganized according to the semantic requirement of the question, moving from general candidate information towards more discriminative evidence for prediction.

Question-conditioned Evidence Reasoning. The influence of question semantics becomes clearer when different questions are asked about the same surgical frame, as illustrated in Fig. 4. Since the visual content remains unchanged, differences in the evidence maps primarily reflect how the reasoning process responds to different semantic requirements. The early-hop responses share several visually prominent regions, but their subsequent evolution becomes increasingly question dependent. A question about the operated organ maintains evidence over the anatomical region, whereas questions about instrument identity, state, or location redirect the reasoning towards different parts of the same scene. In particular, the relevant evidence can shift from the broader operative field to a specific instrument or to the local region associated with its action. These observations are consistent with the interaction between question-state update and visual-evidence refinement in PER. Rather than applying a fixed attention pattern throughout inference, successive hops allow the evidence used for prediction to change as the question representation evolves. Together with the quantitative results in Table 6, the visualizations suggest that the benefit of multi-hop reasoning arises not merely from repeated processing, but from progressively reorganizing visual information according to what is required by the current question.

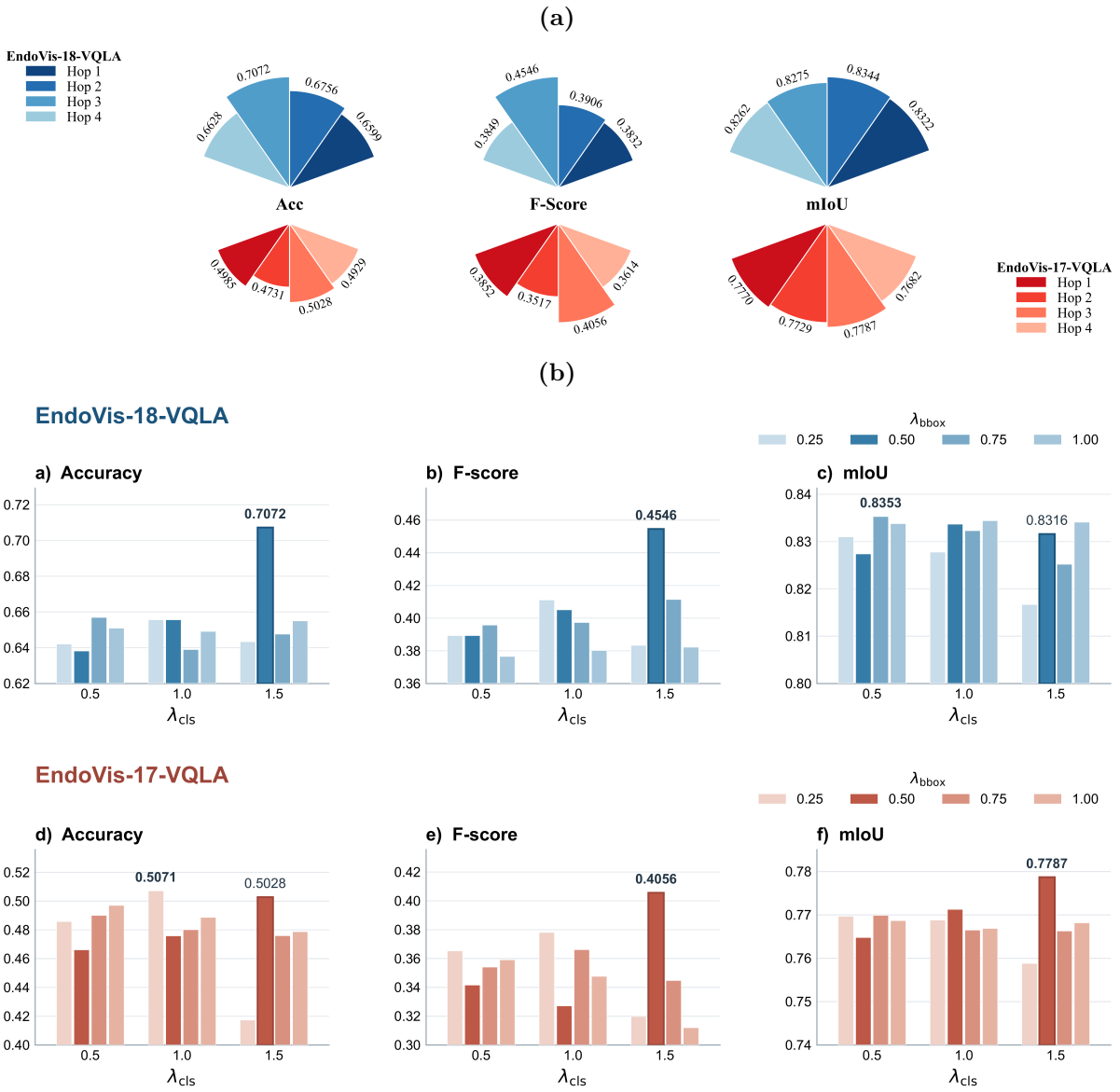


Fig. 5 (a) Comparison of different reasoning depths on EndoVis-18-VQLA and EndoVis-17-VQLA. Acc, F-Score, and mIoU are reported for one to four reasoning hops. (b) Sensitivity analysis of the classification and localization loss weights on EndoVis-18-VQLA and EndoVis-17-VQLA. Accuracy, F-Score, and mIoU are reported under different combinations of λ_{cls} and λ_{bbox}, with the selected setting outlined and panel-wise best values shown in bold. The blue and terracotta palettes distinguish EndoVis-18-VQLA and EndoVis-17-VQLA, respectively.

Loss balancing

To assess the influence of classification and localization supervision on joint optimization, different combinations of λ_{cls} and λ_{bbox} are compared in Fig. 5(b). Increasing the contribution of the bounding-box objective tends to preserve strong localization performance, but does not lead to a consistent improvement in answer discrimination, suggesting that stronger spatial supervision alone cannot guarantee that the localized evidence is sufficiently informative for semantic reasoning. Similarly, increasing the classification contribution does not consistently improve Acc or F-Score, indicating that the two objectives need to be appropriately balanced during joint optimization. Among the evaluated settings, $\lambda_{\text{cls}} = 1.5$ and $\lambda_{\text{bbox}} = 0.5$ provides the most favorable overall trade-off. It achieves the highest Acc and F-Score on EndoVis-18-VQLA (0.7072 and 0.4546), together with the highest F-Score and mIoU on EndoVis-17-VQLA (0.4056 and 0.7787), while maintaining competitive localization performance on EndoVis-18-VQLA. Although individual metrics reach slightly higher values under other weight combinations, no alternative setting provides comparable performance across both answer prediction and localization on the two datasets. These results suggest that effective surgical VQLA depends on balanced supervision between semantic discrimination and spatial grounding rather than favoring either objective alone.

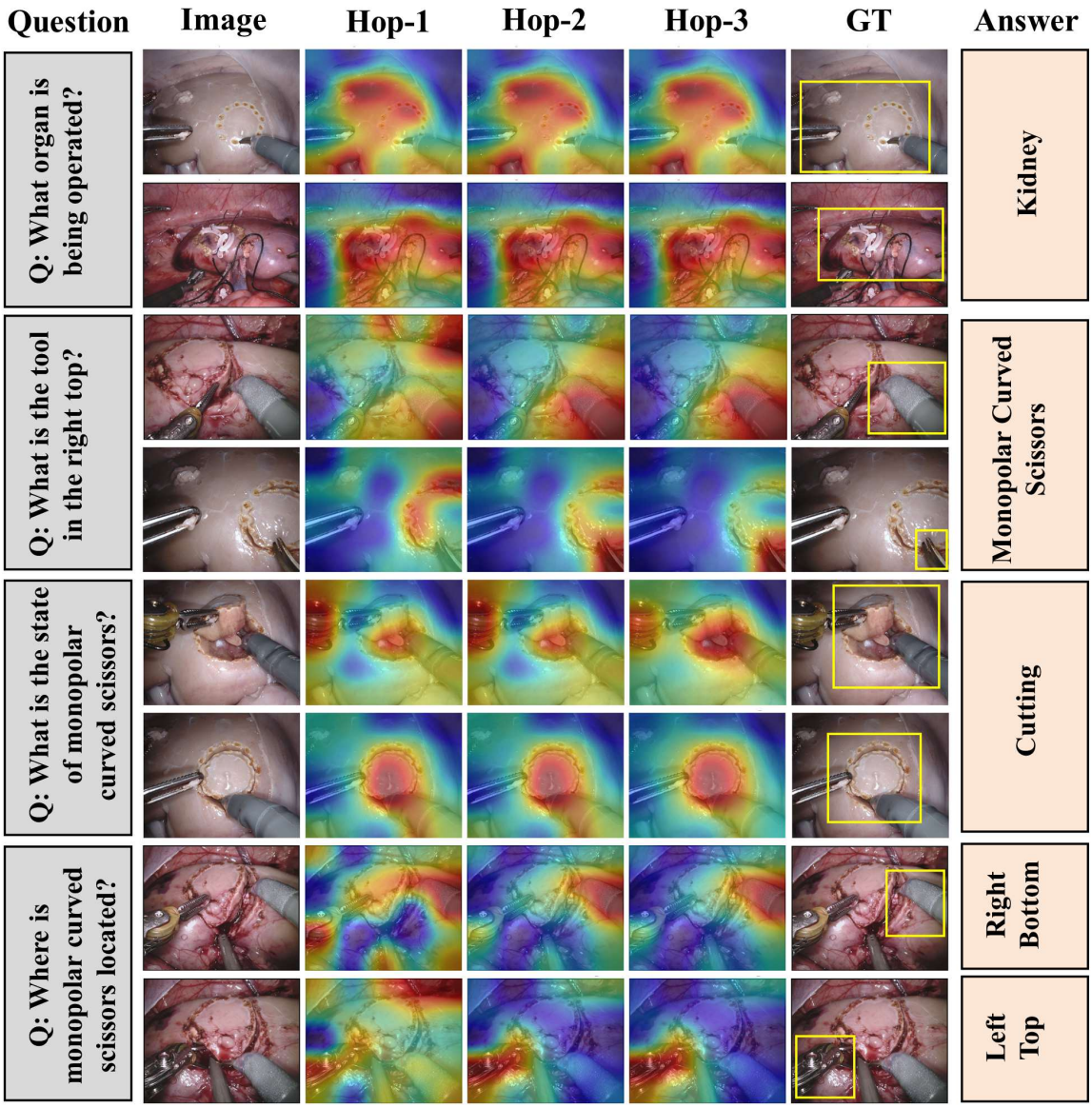


Fig. 6 Visualization of multi-hop evidence evolution for four question types across different surgical scenes. Representative examples of organ, instrument, state, and location questions are shown with the evidence maps from successive reasoning hops, together with the corresponding ground-truth regions and predicted answers.

Discussion

The present results support progressive evidence reasoning as an effective strategy for surgical VQLA. On EndoVis-18-VQLA, SurgPR-Mamba achieves an Acc of 0.7072, an F-Score of 0.4546, and an mIoU of 0.8316, demonstrating improvements in both answer prediction and visual grounding. The contribution of question-guided representation is also reflected in the overall ablation, where introducing QMR increases the F-Score from 0.3619 to 0.4168 and the mIoU from 0.7695 to 0.8305. The internal QMR analysis further shows that incorporating QEA and CMF progressively strengthens question-related evidence modeling, with the complete QMR configuration achieving 0.7072, 0.4546, and 0.8316 on EndoVis-18-VQLA. These results suggest that a substantial part of the difficulty in surgical VQLA lies in establishing a reliable correspondence between the question and the visual evidence required to support its answer. This is particularly important in surgical scenes containing multiple visually similar instruments and localized instrument tissue interactions. Rather than treating all visual information as equally relevant, question-guided representation establishes a spatially grounded evidence state that can be progressively refined during subsequent reasoning.

The behavior of PER further illustrates the complementary roles of evidence localization and semantic interpretation. Introducing the visual update improves localization and cross-dataset performance, while jointly updating the visual evidence and question state further strengthens answer prediction. With both

updates enabled, the model achieves an Acc of 0.7072, an F-Score of 0.4546, and an mIoU of 0.8316 on EndoVis-18-VQLA. On EndoVis-17-VQLA, the corresponding Acc, F-Score, and mIoU reach 0.5028, 0.4056, and 0.7787, respectively. A related pattern is observed in the multi-hop analysis. On EndoVis-18-VQLA, mIoU reaches its highest value of 0.8344 after two reasoning hops, whereas three hops yield the strongest semantic performance, with Acc and F-Score increasing to 0.7072 and 0.4546 while maintaining a comparable mIoU of 0.8316. On EndoVis-17-VQLA, three-hop reasoning achieves the highest values across all three metrics. These results indicate that spatial grounding and semantic reasoning are closely related but do not necessarily evolve synchronously. Once the queried target has been sufficiently localized, additional reasoning can incorporate contextual and relational evidence beyond the most compact target region. Such information is particularly relevant to questions concerning instrument states and interactions, where the answer may depend not only on the instrument itself but also on its relationship with surrounding tissue and the broader surgical context. The evidence visualizations support this interpretation, showing that successive hops progressively reorganize visual responses according to the semantic requirements of different questions rather than simply producing increasingly concentrated attention.

The cross-dataset results further suggest that visual grounding and semantic discrimination exhibit different transfer characteristics. Without using EndoVis-17-VQLA for training, SurgPR-Mamba achieves an Acc of 0.5028, an F-Score of 0.4056, and an mIoU of 0.7787. The relatively strong localization performance indicates that the spatial relationship between a question and its supporting visual region can remain transferable under dataset shift, whereas category-level answer prediction may be more sensitive to variations in instrument appearance, surgical context, and answer distribution. The loss-weight analysis leads to a consistent observation. Individual configurations can achieve slightly higher localization or classification performance on a particular metric, but these improvements do not consistently transfer across both tasks and datasets. In contrast, $\lambda_{\text{cls}} = 1.5$ and $\lambda_{\text{bbox}} = 0.5$ provides the most favorable overall balance, achieving the highest F-Score on both datasets while maintaining strong localization performance. Together, these findings suggest that semantic prediction and spatial grounding should be treated as complementary rather than interchangeable objectives. Reliable surgical VQLA requires not only predicting the correct answer, but also maintaining consistency between that prediction and the visual evidence on which it is based.

Several limitations remain in the current study. First, the evaluation is conducted on EndoVis-18-VQLA and EndoVis-17-VQLA, both of which contain predefined question and answer categories. The ability of progressive evidence reasoning to handle more diverse procedures, instruments, and open-ended surgical questions therefore remains to be established. Second, the current framework operates on individual frames, whereas many surgical states and instrument tissue interactions are inherently temporal. Incorporating temporal evidence across consecutive frames may provide more reliable cues for recognizing dynamic surgical activities. Third, the reasoning depth is fixed for all questions. The multi-hop experiments show that performance improves from one to three hops but declines when the reasoning depth is increased to four, indicating that different questions may not require the same amount of evidence refinement. An adaptive reasoning strategy that determines when sufficient evidence has been acquired would therefore be a meaningful extension. Finally, the evidence maps provide an intuitive view of the evolving reasoning focus but do not constitute a causal explanation of the model decision. Future work should combine progressive reasoning with more rigorous evidence attribution and uncertainty estimation, and evaluate the resulting framework across broader surgical datasets and clinical conditions before considering its use in real-world surgical decision support.

This study presents SurgPR-Mamba for surgical visual question localized answering, where visual evidence is progressively refined according to the semantic requirements of the question rather than remaining fixed after multimodal fusion. Spatially grounded question visual representations establish the initial evidence state, while iterative interaction between visual evidence and question semantics allows relevant cues to be reassessed and refined throughout the reasoning process. SurgPR-Mamba achieves an Acc of 0.7072, an F-Score of 0.4546, and an mIoU of 0.8316 on EndoVis-18-VQLA, with the F-Score and mIoU improving by 13.66 and 3.34 percentage points over Surgical-VQLA++, respectively. The multi-hop analysis further shows that the benefit of progressive reasoning does not arise simply from repeatedly concentrating on a fixed region. Instead, successive reasoning steps reorganize visual evidence according to the information required by the question, enabling semantic prediction to incorporate both localized targets and their surrounding surgical context. This capability is also maintained under cross-dataset evaluation, where the model achieves an Acc of 0.5028, an F-Score of 0.4056, and an mIoU of 0.7787 on EndoVis-17-VQLA. Meanwhile, the selective state-space formulation keeps the model compact, with 174.03M parameters, approximately 5.16× fewer than Surgical-VQLA++. Together, these findings support progressive, question conditioned evidence refinement as an effective means of connecting semantic reasoning with spatial grounding in surgical vision language understanding. This capability could improve

the interpretability of intraoperative decision support and provide a perceptual basis for surgical robots to reason about instruments, anatomy, and instrument tissue interactions in support of more context-aware assistance and future autonomy.

Methods

Datasets

EndoVis-18-VQLA. EndoVis-18-VQLA is constructed from the 2018 MICCAI Endoscopic Vision Challenge dataset [38] and contains video sequences collected from 14 robotic surgical procedures. The VQLA annotations are publicly provided by [8]. In this study, the expanded version introduced in [9] is adopted, where the number of question-answer (QA) pairs is increased from 9,014 to 12,741 for training and from 2,769 to 3,820 for evaluation. Following the data protocol of previous VQLA studies [9, 10], the dataset contains 1,560 training frames and 436 evaluation frames. The same data partition is maintained throughout all experiments to ensure a consistent comparison with existing methods.

As illustrated in Fig. 7, EndoVis-18-VQLA exhibits a pronounced long-tailed distribution at both the question-type and category levels. Organ-related questions are relatively underrepresented, while action, instrument, and location annotations are further dominated by a few frequent classes. This hierarchical imbalance makes minority categories more difficult to recognize and increases the challenge of learning robust semantic and spatial representations from surgical scenes.

EndoVis-17-VQLA. EndoVis-17-VQLA is derived from the 2017 MICCAI Endoscopic Vision Challenge dataset [39], which contains 10 robotic surgery video sequences, with VQLA annotations provided by [8]. Following the established cross-dataset evaluation protocol [9, 17], EndoVis-17-VQLA is used exclusively for external validation to evaluate the generalization ability of the model on unseen surgical scenes. The expanded version of the dataset contains 708 QA pairs, compared with 472 pairs in the original setting [9]. No samples from EndoVis-17-VQLA are involved in model training, allowing the cross-dataset evaluation to directly assess the robustness and transferability of the learned representations.

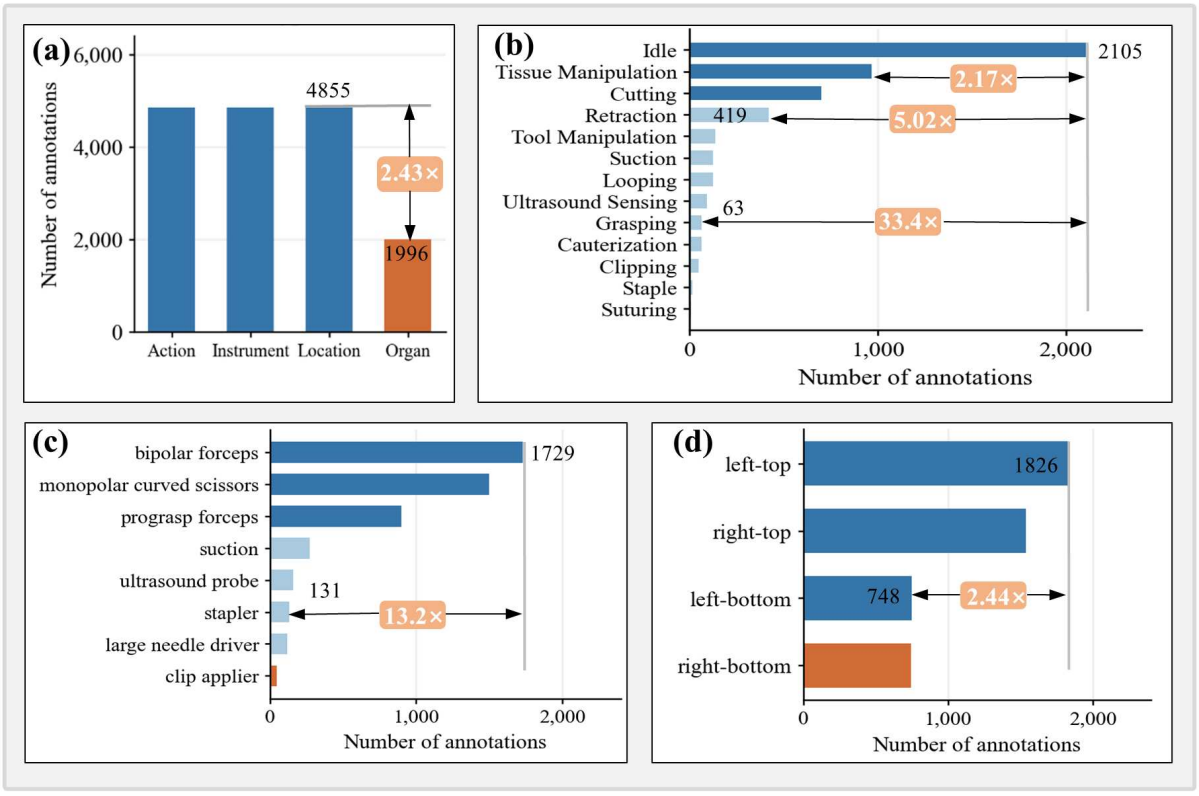


Fig. 7 Annotation statistics of the expanded EndoVis-18-VQLA dataset. (a) Distribution of question types. (b) Long-tailed distribution of action classes. (c) Long-tailed distribution of surgical instrument classes. (d) Distribution of spatial location classes. The dataset exhibits substantial intra-category imbalance, particularly for action and instrument annotations.

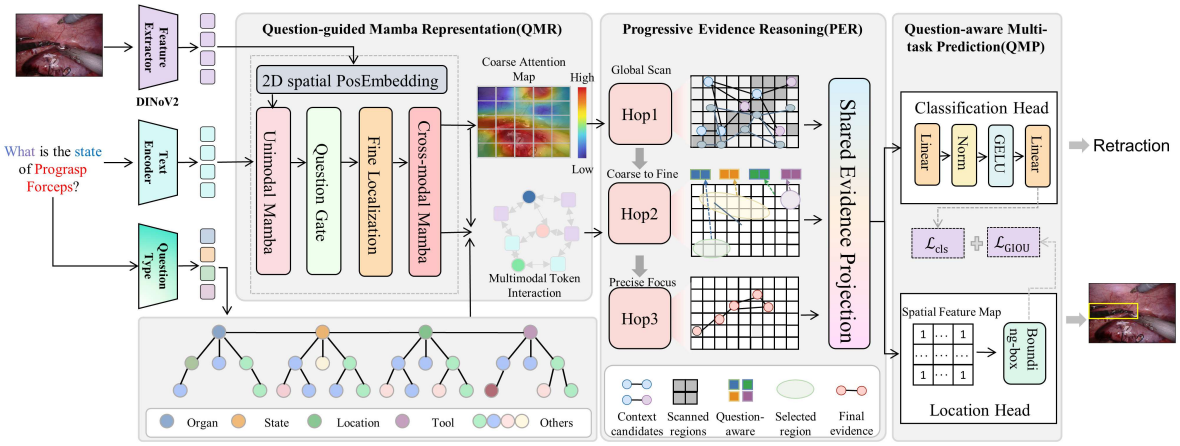


Fig. 8 Overall architecture of SurgPR-Mamba. The framework consists of three main components: Question-guided Mamba Representation (QMR), Progressive Evidence Reasoning (PER), and Question-aware Multi-task Prediction (QMP). QMR integrates spatially encoded visual features, question semantics, and question-type information to construct a question-aware multimodal representation. PER progressively updates and refines the visual evidence through multi-hop reasoning, producing a shared evidence representation from coarse contextual information to fine-grained target-related cues. QMP uses the refined evidence for joint answer classification and bounding-box localization under the classification and GIoU objectives.

Implementation details

All experiments are implemented using PyTorch and conducted on a single NVIDIA GeForce RTX 4090 GPU. DINOv2-Small is adopted as the pretrained visual feature extractor. Each surgical image is represented by 5×5 visual tokens with a feature dimension of 384, which are projected into the multimodal embedding space before being fed into the proposed framework. A dropout rate of 0.1 is applied during training. The model is trained for 80 epochs with a batch size of 64 and an initial learning rate of 8×10^{-6}. To alleviate the long-tailed distribution of answer categories, a weighted sampling strategy is employed, where the sampling weight of each category is inversely proportional to the square root of its sample frequency.

Evaluation metrics

Three metrics are adopted to evaluate answer prediction and visual localization. Accuracy (Acc) [40] measures the overall correctness of predicted answers, while F-Score [40] reflects classification performance by jointly considering precision and recall. Mean Intersection over Union (mIoU) [41] is used to quantify the spatial agreement between predicted and ground-truth bounding boxes. These metrics jointly characterize the semantic accuracy and localization quality of the model.

Overall framework

The overall architecture of SurgPR-Mamba is illustrated in Fig. 8. The framework is designed around the idea that the visual evidence required to answer a surgical question should remain spatially grounded while being progressively adapted to the evolving semantic context of the question. To this end, SurgPR-Mamba integrates Question-guided Mamba Representation (QMR), Progressive Evidence Reasoning (PER), and Question-aware Multi-task Prediction (QMP) within a unified reasoning framework. QMR establishes the initial question-aware representation by preserving spatial structure, emphasizing question-relevant visual information, and modeling contextual dependencies across visual and linguistic features. Rather than treating this representation as fixed, PER further allows the visual evidence and question state to evolve through recurrent interaction, so that different reasoning steps can emphasize complementary cues associated with the surgical context, queried target, and local instrument tissue relationships. The resulting evidence is finally exploited by QMP to support answer discrimination and visual localization within a shared question-conditioned representation. Through this design, SurgPR-Mamba couples multimodal representation learning with progressive evidence refinement, enabling semantic reasoning and spatial grounding to remain closely coordinated throughout surgical VQLA.

Question-guided multimodal representation

The QMR module is designed to establish a strong connection between the meaning of the question and the spatial visual information in complex surgical scenes. As illustrated in Fig. 8, both visual and linguistic

features are encoded using a spatially aware Mamba architecture, which preserves the spatial layout of surgical areas as well as the contextual relationships within each modality. The encoded question semantics then guide the alignment of visual evidence, enabling the representation to focus on regions relevant to the specific question rather than treating all visual data equally. Subsequently, cross-modal Mamba interaction spreads and integrates the aligned information across both visual and linguistic sequences, producing a question-aware multimodal representation that incorporates spatial grounding and semantic relevance.

Spatial-aware Mamba encoding

Given a surgical image I and its corresponding question Q, a pretrained DINOv2 visual encoder is employed to extract visual features $\mathbf{V} \in \mathbb{R}^{N \times d_v}$, while the linguistic encoder maps the question into textual features $\mathbf{T} \in \mathbb{R}^{L \times d_t}$, where N and L denote the numbers of visual and linguistic tokens, respectively, and d is the embedding dimension. To retain the spatial organization of the surgical scene, a learnable two-dimensional positional representation is introduced into the visual tokens. For a visual token located at the i-th row and j-th column, its spatial embedding is defined as

$$\mathbf{p}_{i,j} = [\phi_r(r_i) \| \phi_c(c_j)] \,, \tag{1}$$

where r_i and c_j are the normalized row and column coordinates, $\phi_r(\cdot)$ and $\phi_c(\cdot)$ denote learnable linear projections, and $\|$ represents concatenation. The spatial embeddings of all visual tokens form $\mathbf{P} \in \mathbb{R}^{N \times d}$ and are incorporated into the original visual representation through

$$\mathbf{V}^s = \mathrm{LN}\left(\mathbf{V} + \mathbf{P}\right), \tag{2}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization. This operation preserves the visual semantics while explicitly encoding the two-dimensional spatial relationships among surgical regions.

The spatially enhanced visual tokens and linguistic tokens are subsequently modeled by modality-specific Mamba encoders:

$$\mathbf{V}^m = \mathcal{M}_v\left(\mathbf{V}^s\right), \qquad \mathbf{T}^m = \mathcal{M}_q\left(\mathbf{T}\right), \tag{3}$$

where $\mathcal{M}_v(\cdot)$ and $\mathcal{M}_q(\cdot)$ denote the visual and linguistic Mamba encoders, respectively. The selective state-space modeling captures contextual dependencies within each modality while maintaining the spatial structure of the visual sequence. The resulting representations $\mathbf{V}^m$ and $\mathbf{T}^m$ provide spatially structured visual context and contextualized question semantics for question-guided evidence alignment.

Question-guided evidence alignment

The encoded visual representation contains information from the entire surgical scene, while only a subset of visual regions is relevant to a specific question. To establish an explicit correspondence between question semantics and visual evidence, a question-guided evidence alignment mechanism is introduced to adaptively emphasize informative regions. A global question representation $\mathbf{q} \in \mathbb{R}^d$ is derived from the contextualized linguistic features $\mathbf{T}^m$ and used to estimate the relevance of each visual token:

$$\mathbf{G} = \sigma\left(\mathbf{W}_g \tanh\left(\mathbf{W}_v \mathbf{V}^m + \mathbf{W}_q \mathbf{q}\right)\right), \tag{4}$$

where $\mathbf{W}_v$, $\mathbf{W}_q$, and $\mathbf{W}_g$ are learnable projections, and $\sigma(\cdot)$ denotes the sigmoid function. The relevance gate $\mathbf{G}$ modulates the visual representation according to the semantic intent of the question, yielding

$$\widetilde{\mathbf{V}} = \mathrm{LN}\left(\mathbf{V}^m + \mathbf{G} \odot \mathbf{V}^m\right), \tag{5}$$

where $\odot$ denotes element-wise multiplication. The residual formulation retains the original scene context while strengthening visual evidence associated with the question.

Beyond global question guidance, fine-grained token-level alignment is further introduced to capture local correspondence between visual regions and linguistic cues. The cross-modal relevance matrix is computed as

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\left(\widetilde{\mathbf{V}}\mathbf{W}_q\right)\left(\mathbf{T}^m\mathbf{W}_K\right)^\top}{\sqrt{d}}\right), \tag{6}$$

where $\mathbf{W}_q$ and $\mathbf{W}_K$ are learnable projection matrices. Based on the learned correspondence, question-conditioned linguistic evidence is aggregated for each visual token as

$$\mathbf{E} = \mathbf{A}\left(\mathbf{T}^m\mathbf{W}_v\right), \tag{7}$$

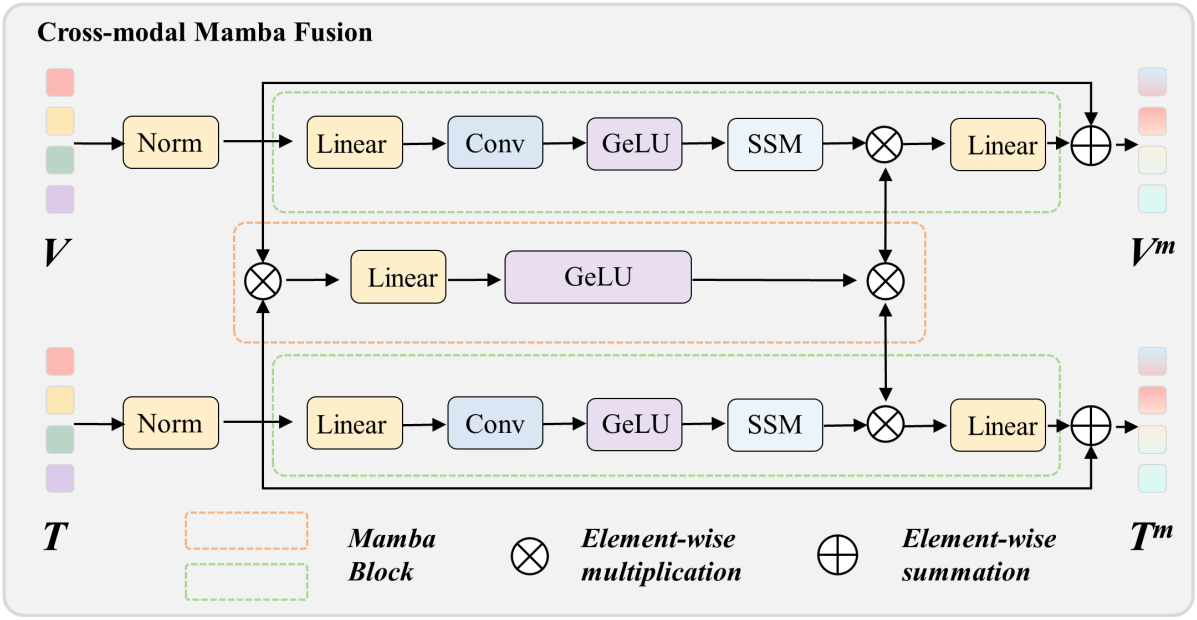


Fig. 9 Illustration of the Cross-modal Mamba Fusion module. Parallel Mamba branches model the contextual dependencies of visual and textual features, while cross-modal interaction provides complementary information to both streams. The interacted features are then integrated through residual connections to obtain the updated visual and textual representations $\mathbf{V}^m$ and $\mathbf{T}^m$.

where $\mathbf{W}_v$ denotes the value projection. In this way, each visual region is associated with the linguistic cues most relevant to its semantic content. The aligned representation $\mathbf{E}$ complements the question-enhanced visual features $\widetilde{\mathbf{V}}$, providing fine-grained visual-language evidence for subsequent cross-modal Mamba fusion.

Cross-modal Mamba fusion

The aligned visual and linguistic evidence provides localized cross-modal correspondence, but effective reasoning further requires contextual interaction between the two modalities, as illustrated in Fig. 9. To this end, a cross-modal Mamba mechanism is employed to propagate question-related evidence across the visual sequence. The question-aligned evidence $\mathbf{E}$ is integrated with the enhanced visual representation $\widetilde{\mathbf{V}}$ and modeled through a cross-modal Mamba block:

$$\mathbf{Z}^c = \mathcal{M}_c\left(\widetilde{\mathbf{V}}, \mathbf{E}\right), \tag{8}$$

where $\mathcal{M}_c(\cdot)$ denotes the cross-modal Mamba operation. Through selective state-space modeling, question-related information is propagated across visual tokens, allowing local evidence to be interpreted within the broader surgical context.

Rather than directly replacing the original visual representation with the cross-modal output, a token-level gating mechanism is introduced to adaptively balance visual context and question-conditioned evidence. The fusion gate is computed as

$$\boldsymbol{\gamma} = \sigma\left(\mathbf{W}_f\left[\widetilde{\mathbf{V}} \| \mathbf{Z}^c\right] + \mathbf{b}f\right), \tag{9}$$

where $\mathbf{W}_f$ and $\mathbf{b}_f$ are learnable parameters. The final multimodal representation is obtained by

$$\mathbf{Z}^0 = \boldsymbol{\gamma} \odot \mathbf{Z}^c + (1 - \boldsymbol{\gamma}) \odot \widetilde{\mathbf{V}}. \tag{10}$$

The token-wise gate allows different visual regions to retain complementary visual context or incorporate question-conditioned information according to their relevance. Consequently, $\mathbf{Z}^0$ encodes spatial structure, question semantics, and localized cross-modal evidence within a unified representation, and serves as the initial evidence state for the subsequent Progressive Evidence Reasoning module.

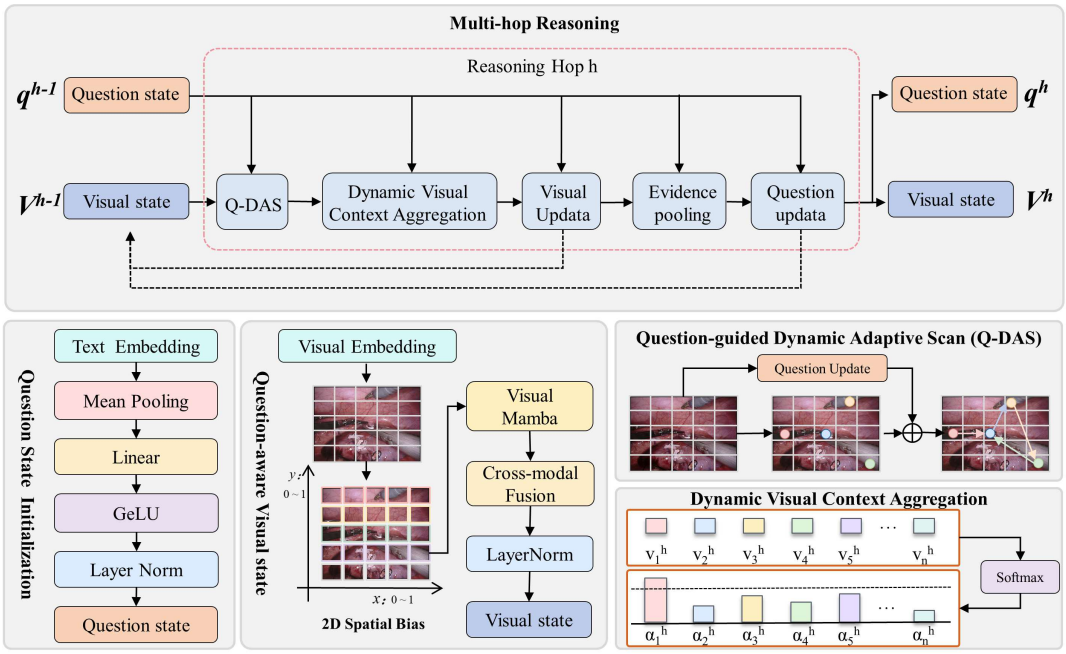


Fig. 10 Overview of the Multi-hop Reasoning mechanism. The framework maintains coupled question and visual states throughout reasoning, where question semantics guide adaptive visual context modeling and evidence extraction, and the acquired evidence is fed back to update the reasoning state. This recurrent interaction enables visual evidence and question semantics to be jointly refined across successive hops.

Progressive evidence reasoning

The question-aware multimodal representation obtained from QMR establishes the initial correspondence between visual content and question semantics, yet the evidence required for prediction may involve different levels of contextual and relational information. The PER module therefore organizes the multimodal representation into an iterative reasoning process, where visual evidence and question states are progressively refined across multiple reasoning hops. At each hop, the current question state guides the update of visual evidence, from which informative evidence is extracted to support the current reasoning step [27]. The extracted evidence is further incorporated into the question representation, allowing the updated question state to guide the next hop. Such recurrent interaction forms a closed reasoning loop between visual evidence and question semantics, enabling the reasoning focus to evolve from broad contextual cues toward more specific and discriminative evidence. After H reasoning hops, the accumulated evidence provides the refined multimodal representation for answer prediction and visual localization.

Visual evidence update

As shown in Fig. 10, at the h-th reasoning hop, the visual evidence is dynamically updated according to the current question state, allowing the model to adjust its reasoning focus as the inference proceeds [42]. Given the visual representation $\mathbf{V}^{h-1}$ and question state $\mathbf{q}^{h-1}$ from the previous hop, a question-conditioned relevance gate is constructed to estimate the importance of each visual token:

$$\mathbf{G}^{h} = \sigma\left(\mathbf{W}_{g}\left[\mathbf{V}^{h-1} \| \mathbf{q}^{h-1}\right] + \mathbf{b}_{g}\right), \tag{11}$$

where $\mathbf{W}_{g}$ and $\mathbf{b}_{g}$ are learnable parameters, and $\sigma(\cdot)$ denotes the sigmoid function. The question state is broadcast along the visual sequence to establish token-wise semantic guidance.

To propagate contextual information according to the current reasoning state, a Question-guided Dynamic Adaptive Scan (Q-DAS) is introduced, as illustrated in Fig. 10. Rather than applying an identical update to all visual tokens, Q-DAS adaptively balances the original visual evidence and the question-conditioned contextual representation:

$$\widetilde{\mathbf{V}}^{h} = \mathbf{G}^{h} \odot \mathbf{V}^{h-1} + \left(1 - \mathbf{G}^{h}\right) \odot \mathcal{S}\left(\mathbf{V}^{h-1}, \mathbf{q}^{h-1}\right), \tag{12}$$

where $\mathcal{S}(\cdot)$ denotes the question-conditioned adaptive scanning operation and $\odot$ represents element-wise multiplication. The updated representation is further modeled by the Mamba state-space operation:

$$\mathbf{V}^h = \mathrm{LN}\left(\widetilde{\mathbf{V}}^h + \mathcal{M}_h\left(\widetilde{\mathbf{V}}^h\right)\right), \tag{13}$$

where $\mathcal{M}_h(\cdot)$ denotes the Mamba reasoning operation at the h-th hop. In this manner, the visual state is continuously reorganized under question guidance, allowing different reasoning hops to emphasize complementary evidence rather than repeatedly operating on a fixed visual representation. The resulting $\mathbf{V}^h$ is subsequently used to extract the evidence required by the current reasoning step.

Multi-hop reasoning

The updated visual state $\mathbf{V}^h$ contains contextual information accumulated at the h-th reasoning hop, while its contribution to the current reasoning step varies across visual regions. The current question state $\mathbf{q}^{h-1}$ is therefore used to evaluate the relevance of individual visual tokens and extract the evidence required for the current inference. The question-conditioned relevance score is computed as

$$e_i^h = \mathbf{w}_e^\top \tanh\left(\mathbf{W}_v^e \mathbf{v}_i^h + \mathbf{W}_q^e \mathbf{q}^{h-1}\right), \tag{14}$$

where $\mathbf{v}_i^h$ denotes the i-th visual token at the h-th hop, and $\mathbf{W}_v^e$, $\mathbf{W}_q^e$, and $\mathbf{w}_e$ are learnable parameters. The normalized evidence distribution over the visual tokens is obtained by

$$\alpha_i^h = \frac{\exp(e_i^h)}{\sum_{j=1}^{N} \exp(e_j^h)}. \tag{15}$$

The evidence representation of the current reasoning hop is then aggregated as

$$\mathbf{e}^h = \sum_{i=1}^{N} \alpha_i^h \mathbf{v}_i^h. \tag{16}$$

The resulting $\mathbf{e}^h$ summarizes the visual information that is most compatible with the current question state. Since both $\mathbf{V}^h$ and $\mathbf{q}^{h-1}$ evolve throughout the reasoning process, the evidence distribution is re-estimated at each hop rather than shared across all reasoning steps. This allows successive hops to attend to different but complementary visual cues, providing an evolving evidence representation for updating the question state.

Question state update

The evidence extracted at each reasoning hop provides new visual information for interpreting the question. Instead of keeping the question representation fixed throughout the reasoning process, the question state is dynamically updated with the evidence obtained from the current hop. Given the previous question state $\mathbf{q}^{h-1}$ and the extracted evidence $\mathbf{e}^h$, an evidence-conditioned question representation is constructed as

$$\widehat{\mathbf{q}}^h = \phi_q\left(\left[\mathbf{q}^{h-1} \| \mathbf{e}^h\right]\right), \tag{17}$$

where $\phi_q(\cdot)$ denotes a learnable transformation and $\|$ represents feature concatenation. A gating function is introduced to control the contribution of newly acquired evidence to the evolving question state:

$$\boldsymbol{\beta}^h = \sigma\left(\mathbf{W}_\beta\left[\mathbf{q}^{h-1} \| \mathbf{e}^h\right] + \mathbf{b}_\beta\right), \tag{18}$$

where $\mathbf{W}_\beta$ and $\mathbf{b}_\beta$ are learnable parameters. The question state is updated as

$$\mathbf{q}^h = \mathrm{LN}\left(\mathbf{q}^{h-1} + \boldsymbol{\beta}^h \odot \widehat{\mathbf{q}}^h\right). \tag{19}$$

The updated state $\mathbf{q}^h$ retains the original semantic intent of the question while incorporating the visual evidence acquired at the current reasoning hop. It subsequently guides the visual evidence update of the next hop, allowing question interpretation and visual evidence to evolve jointly throughout the reasoning process. The interaction among visual evidence update, evidence extraction, and question state update forms an iterative reasoning cycle, through which the model progressively concentrates on evidence that is increasingly relevant to the final prediction.

Question-aware multi-task prediction

The representation obtained after progressive evidence reasoning is used to jointly perform answer prediction and visual localization. The accumulated visual evidence is integrated with the refined question representation to preserve question-specific semantics during prediction. Question-type information is further embedded into the representation, providing complementary semantic cues for distinguishing different reasoning objectives in surgical scenes. The resulting question-aware representation $\mathbf{z}$ is formulated as

$$\mathbf{z} = \phi_p\left(\left[\mathbf{e}^H \| \mathbf{q}^H \| \mathbf{t}_q\right]\right), \tag{20}$$

where $\mathbf{e}^H$ and $\mathbf{q}^H$ denote the visual evidence and question state obtained after the final reasoning hop, respectively, $\mathbf{t}_q$ represents the question-type embedding, and $\phi_p(\cdot)$ denotes a learnable projection. The shared representation is subsequently fed into two task-specific prediction heads:

$$\hat{\mathbf{y}} = \mathcal{H}_{\text{ans}}(\mathbf{z}), \qquad \hat{\mathbf{b}} = \mathcal{H}_{\text{loc}}(\mathbf{z}), \tag{21}$$

where $\mathcal{H}_{\text{ans}}(\cdot)$ and $\mathcal{H}_{\text{loc}}(\cdot)$ denote the answer classification and bounding-box regression heads, respectively. In this manner, the final prediction is jointly conditioned on the progressively refined visual evidence and question semantics, allowing answer recognition and evidence localization to share a consistent multimodal representation.

Loss function

The proposed framework is trained in an end-to-end manner by jointly optimizing answer prediction and visual localization. For answer prediction, the classification objective is defined using the cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_c \log \hat{p}_c, \tag{22}$$

where C denotes the number of answer categories, y_c is the ground-truth label, and $\hat{p}_c$ represents the predicted probability for the c-th category.

For visual localization, the Generalized Intersection over Union (GIoU) loss is adopted to measure the spatial discrepancy between the predicted bounding box $\hat{\mathbf{b}}$ and the ground-truth bounding box $\mathbf{b}$:

$$\mathcal{L}_{\text{bbox}} = 1 - \text{GIoU}\left(\hat{\mathbf{b}}, \mathbf{b}\right). \tag{23}$$

The overall training objective is formulated as a weighted combination of the two tasks:

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{bbox}}\mathcal{L}_{\text{bbox}}, \tag{24}$$

where λ_{cls} and λ_{bbox} control the contributions of answer prediction and visual localization, respectively. This joint optimization encourages the model to learn discriminative semantic representations for answer prediction while maintaining spatial consistency with the corresponding visual evidence.

Data availability

The datasets used in this study are publicly available and can be accessed at (https://drive.google.com/file/d/1-FXOdhD3uw55ATDgI1wPEe-txyuCiP2E/view). Further details regarding the datasets and their use in this study are provided in the Datasets subsection of the Methods section.

Code availability

The source code for SurgPR-Mamba will be made publicly available upon acceptance of the manuscript at https://github.com/Roseliangg/SurgPR-Mamba.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2025YFC2426000); in part by the Shenzhen Key Industry Research and Development Program

(ZDCY20251230093010012); in part by the National Natural Science Foundation of China under Grant 62473053; in part by the Beijing Natural Science Foundation (No.L232004); in part by the Beijing Natural Science Foundation (No.L252173).

Author contributions

Lianliang Li: Conceptualization, Methodology, Formal analysis, Writing- Original draft; Xue Li: Data curation, Resources; Weiyan Kong:Data curation; Zhaoyun Wu:Data curation; Changsheng Li: Reviewing, Funding acquisition; Xingguang Duan: Reviewing, Supervision.

Competing interests

The authors declare no competing interests.

References

[1] Schmidgall, S., Kim, J. W., Kuntz, A., Ghazi, A. E. & Krieger, A. General-purpose foundation models for increased autonomy in robot-assisted surgery. *Nature Machine Intelligence* **6**, 1275–1283 (2024).

[2] Peng, J. *et al.* Optical metasurfaces for general vision processing on the edge. *Nature* **654**, 917–925 (2026).

[3] Long, Y. *et al.* Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. *Science Robotics* **10**, eadt3093 (2025).

[4] Li, X., Chen, W., Duan, X., Gu, X. & Li, C. Collaborative surgical instrument segmentation for monocular depth estimation in minimally invasive surgery. *Medical Image Analysis* 103765 (2025).

[5] He, R. *et al. Pitvqa: Image-grounded text embedding llm for visual question answering in pituitary surgery. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 488–498 (Springer, 2024).

[6] Seenivasan, L., Islam, M., Krishna, A. K. & Ren, H. *Surgical-vqa: Visual question answering in surgical scenes using transformer. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 33–43 (Springer, 2022).

[7] Wang, G. *et al.* Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. *arXiv preprint arXiv:2405.10948* (2024).

[8] Bai, L., Islam, M., Seenivasan, L. & Ren, H. *Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. 2023 IEEE International Conference on Robotics and Automation (ICRA)*, 6859–6865 (IEEE, 2023).

[9] Bai, L. *et al.* Surgical-vqla++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery. *Information Fusion* **113**, 102602 (2025).

[10] 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* **29**, 9027–9040 (2025).

[11] Bai, L., Islam, M. & Ren, H. *Cat-vil: Co-attention gated vision-language embedding for visual question localized-answering in robotic surgery. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 397–407 (Springer, 2023).

[12] Wang, Z. *et al.* Parallel multi-attention and gated fusion for visual question localized answering in surgical scenes. *IEEE Journal of Biomedical and Health Informatics* (2025).

[13] Choi, T.-M. *et al.* Surgmllmbench: A multimodal large language model benchmark dataset for surgical scene understanding. *arXiv preprint arXiv:2511.21339* (2025).

[14] Zhu, L. *et al.* Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417* (2024).

[15] Liu, Y. *et al.* Vmamba: Visual state space model. *Advances in neural information processing systems* **37**, 103031–103063 (2024).

[16] Jin, Q. *et al.* Cmi-mtl: Cross-mamba interaction based multi-task learning for medical visual question answering. *arXiv preprint arXiv:2511.01357* (2025).

[17] Hao, P. *et al. Surgical-mamballm: Mamba2-enhanced multimodal large language model for vqla in robotic surgery. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 573–583 (Springer, 2025).

[18] Quinn, T. *et al.* Influence of surgical expertise on visual gaze patterns while viewing a robotic adrenalectomy. *Journal of Surgical Education* **83**, 103771 (2026).

[19] Liu, K. *et al.* Objective assessment of visual attention in orthognathic surgery training based on eye tracking. *Journal of Cranio-Maxillofacial Surgery* **52**, 65–70 (2024).

[20] Nwoye, C. I. *et al.* Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis* **78**, 102433 (2022).

[21] Yuan, K. *et al.* Advancing surgical vqa with scene graph knowledge. *International journal of computer assisted radiology and surgery* **19**, 1409–1417 (2024).

[22] Li, P. *et al.* Surgical video workflow analysis via visual-language learning. *npj Health Systems* **2**, 5 (2025).

[23] Xiao, J., Yao, A., Li, Y. & Chua, T.-S. *Can i trust your answer? visually grounded video question answering. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13204–13214 (IEEE, 2024).

[24] Lu, Y., Song, Y., Wang, W., Torresani, L. & Nagarajan, T. *Vited: Video temporal evidence distillation. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8501–8511 (IEEE, 2025).

[25] Zhu, Z. *et al. Alignment before awareness: Towards visual question localized-answering in robotic surgery via optimal transport and answer semantics. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 711–721 (2024).

[26] Hao, P. *et al.* Surgery-r1: Advancing surgical-vqla with reasoning multimodal large language model via reinforcement learning. *arXiv preprint arXiv:2506.19469* (2025).

[27] Xu, Z. *et al.* A question-guided multi-hop reasoning graph network for visual question answering. *Information Processing & Management* **60**, 103207 (2023).

[28] Zhang, Y. *et al. Groundhog grounding large language models to holistic segmentation. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14227–14238 (IEEE, 2024).

[29] Chen, B. *et al. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14455–14465 (2024).

[30] Shin, J. *et al. Towards holistic surgical scene graph. International Conference on Medical Image Computing and Computer-Assisted Intervention*, 617–626 (Springer, 2025).

[31] Yu, Z., Yu, J., Cui, Y., Tao, D. & Tian, Q. *Deep modular co-attention networks for visual question answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6274–6283 (IEEE, 2019).

[32] Touvron, H. *et al. Training data-efficient image transformers & distillation through attention. International conference on machine learning*, 10347–10357 (PMLR, 2021).

[33] Ben-Younes, H., Cadene, R., Cord, M. & Thome, N. *Mutan: Multimodal tucker fusion for visual question answering. 2017 IEEE international conference on computer vision (ICCV)*, 2631–2639 (IEEE, 2017).

[34] 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**, 5947–5959 (2018).

[35] Ben-Younes, H., Cadene, R., Thome, N. & Cord, M. *Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, 8102–8109 (2019).

[36] Seenivasan, L., Islam, M., Kannan, G. & Ren, H. *Surgicalgpt: end-to-end language-vision gpt for visual question answering in surgery. International conference on medical image computing and computer-assisted intervention*, 281–290 (Springer, 2023).

[37] Nwoye, C. I. *et al.* Cholectriplet2022: Show me a tool and tell me the triplet—an endoscopic vision challenge for surgical action triplet detection. *Medical Image Analysis* **89**, 102888 (2023).

[38] Allan, M. *et al.* 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190* (2020).

[39] Allan, M. *et al.* 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426* (2019).

[40] Hossin, M. & Sulaiman, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process* **5**, 1 (2015).

[41] Rezatofighi, H. *et al.* Generalized intersection over union: A metric and a loss for bounding box regression. *arXiv preprint arXiv:1902.09630* (2019).

[42] Sun, C., Chen, J., Ma, M. & Guo, L. Multi-perspective knowledge-aware reinforcement learning framework for multi-hop reasoning with temporal constraints. *Information Processing & Management* **64**, 104982 (2026).