



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Towards Unified Surgical Scene Understanding: Bridging Reasoning and Grounding via MLLMs

Jincai Huang^{1,2†}, Shihao Zou^{3†}, Yuchen Guo⁴, Jingjing Li⁵, Wei Ji⁶, Kai Wang⁷, Shanshan Wang³, and Weixin Si^{2✉}

¹Southern University of Science and Technology

²Shenzhen University of Advanced Technology

³Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

⁴Northwestern University ⁵University of Alberta

⁶Yale University ⁷Nanfeng Hospital

siweixin@suat-sz.edu.cn

† These authors contributed equally.

Abstract. Surgical scene understanding is essential for AI-assisted intervention, yet real-world clinical applications require a holistic understanding of procedural context, semantic reasoning, and precise visual grounding. However, existing approaches typically address these components in isolation, leading to fragmented representations and limited semantic consistency. To address this limitation, we propose SurgMLLM, a unified surgical scene understanding framework that bridges high-level reasoning and low-level visual grounding. Given surgical videos, SurgMLLM fine-tunes a multi-modal large language model (MLLM) to support structured interpretability reasoning over phases, instrument-verb-target (IVT) triplets, and triplet-entity segmentation tokens. These tokens are then temporally aggregated and serve as prompts for a segmentation network, enabling accurate pixel-wise grounding of instruments and targets. The framework is trained end-to-end with a unified objective that couples language-based reasoning supervision with visual grounding losses. To facilitate unified evaluation, we introduce CholecT45-Scene, extending CholecT45 dataset with 64,299 frames of pixel-level mask annotations for instruments and targets, aligned with existing triplet labels. Extensive experiments show that SurgMLLM improves the primary triplet recognition metric AP_{IVT} from 40.7% to 46.0% and consistently outperforms prior methods in phase recognition and segmentation. These results highlight the effectiveness of unified reasoning-and-grounding for reliable, context-aware surgical assistance. Our code and dataset are available at <https://github.com/Su-Llo/SurgMLLM>.

Keywords: Surgical Scene Understanding · Surgical Triplet Recognition · Multi-Modal Large Language Model

1 Introduction

Minimally invasive and robot-assisted surgery [17] relies primarily on endoscopic videos as the main source of intra-operative information. To enable context-aware

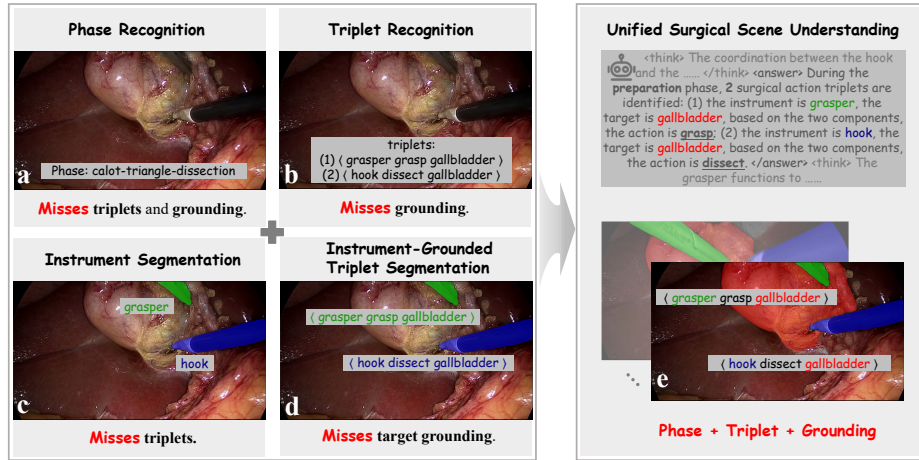


Fig. 1. Representative tasks in surgical scene understanding. (a) Phase recognition [13] provides only high-level workflow state prediction. (b) Triplet recognition [10] identifies only high-level *IVT* interactions without visual grounding. (c) Instrument segmentation [14] performs only pixel-wise semantic recognition of surgical instruments. (d) Recent instrument-grounded triplet segmentation [1] predicts triplet entities while segmenting instruments at the pixel level, but neglecting phase recognition and target grounding. (e) Our framework bridges high-level phase and triplet reasoning with low-level triplet entity grounding via MLLM, enabling comprehensive and coherent surgical scene understanding for robust computer-assisted intervention.

computer-assisted intervention, an intelligent system should not only recognize workflow progression over time, but also reason about fine-grained instrument–tissue interactions and ground these semantics to image regions for interpretable decision support [20]. However, achieving coherent reasoning across temporal, relational, and spatial dimensions remains a fundamental challenge.

Prior research has advanced key components of surgical scene understanding, yet largely in isolation. For temporal workflow modeling, phase recognition evolved from convolution-recurrent pipelines to transformer-based long-range modeling (*e.g.*, Trans-SVNet [7] and SurgFormer [13]). Complementarily, triplet recognition models *IVT* interactions [19, 9, 10], while instrument segmentation [15] provides pixel-level localization without interaction reasoning. Recent instrument-grounded triplet segmentation methods [1] begin to couple prediction and grounding, but often treat targets as categorical labels with weak spatial constraints, and typically omit phase recognition or full triplet-entity grounding. As shown in Fig. 1, this fragmented modeling hinders coherent understanding across temporal, relational, and spatial dimensions.

Importantly, *surgical scenes exhibit strong intrinsic priors*: workflow phases constrain plausible *IVT* interactions, while recurring *IVT* patterns provide cues about phase progression. Ignoring these dependencies prevents models from fully leveraging domain knowledge. Moreover, for computer-assisted interven-

tion, high-level predictions alone are insufficient—semantic decisions must be grounded to precise image regions to ensure reliability and verifiability [3]. When attempting to unify these tasks, additional challenges arise, including aligning workflow evolution with interaction semantics, enforcing cross triplet-entity consistency, and maintaining temporally coherent pixel-level grounding.

Meanwhile, MLLM-based grounded reasoning has shown promise in bridging language reasoning and pixel-level representations. Prompt-driven segmentation models such as the SAM family [11, 22], pixel-grounded conversational models [21, 16], and token-based methods such as LISA [12] and Sa2VA [25] have progressively advanced reasoning-guided mask prediction. In the surgical domain, GP-VLS [23], SurgVLM [26], and EndoChat [24] extend vision-language modeling toward surgical knowledge, scene understanding, and dialogue assistance. However, existing generic grounded MLLMs and surgical MLLMs still do not explicitly couple structured surgical semantics (*e.g.*, *IVT* triplets and workflow phases) with temporally consistent pixel-level grounding of both instruments and targets, motivating a unified surgical-specific MLLM framework.

To address these limitations, we propose **SurgMLLM**, a unified framework that integrates phase recognition, triplet reasoning, and triplet-entity grounding within an MLLM-driven architecture. Given surgical videos, SurgMLLM first performs structured scene reasoning through explicit textual analysis, modeling intrinsic correlations between workflow phases and *IVT* entities. The generated entity-aware [SEG] tokens serve as language-conditioned prompts that bridge high-level semantic reasoning and low-level pixel decoding via SAM2 [22]. A temporal fusion mechanism further enforces cross-frame consistency for grounded entities. By jointly optimizing reasoning and segmentation objectives, the model aligns what it infers, predicts, and grounds within a coherent pipeline.

Our contributions are summarized as follows: (1) We introduce a unified MLLM-based framework for comprehensive surgical scene understanding, jointly addressing workflow phase recognition, *IVT* triplet reasoning, and triplet-entity grounding within a single model. (2) We design an explicit bridge from reasoning to grounding via dedicated tokens and temporal prompt fusion, enabling semantically aligned and temporally consistent segmentation. (3) We construct the CholecT45-Scene dataset with 64,299 frames of triplet pixel-level annotations and structured reasoning narratives to support unified supervision. (4) Extensive experiments demonstrate consistent improvements across all three tasks, establishing a strong benchmark for integrated surgical scene understanding.

2 Method

Problem Definition. The overall pipeline is shown in Fig. 2. Given a T -frame surgical video $\mathcal{X} = \{F_t\}_{t=1}^T$ as input, where each frame $F_t \in \mathbb{R}^{H \times W \times 3}$ has height H and width W , the goal is to jointly predict temporal semantics and spatial grounding for each frame. Specifically, for the t -th frame, the semantics include a workflow phase label p_t , a set of surgical triplets $\{(I_{t,n}, V_{t,n}, O_{t,n})\}_{n=1}^N$, and the corresponding pixel-level grounding masks for the instrument $I_{t,n}$ and target

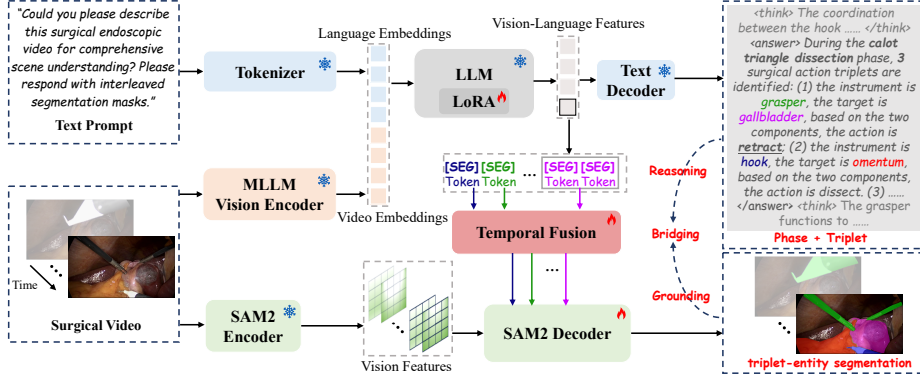


Fig. 2. Overview of SurgMLLM. Given surgical videos, SurgMLLM first performs structured scene reasoning with an MLLM to predict workflow phases and *IVT* triplets, accompanied with triplet-entity [SEG] tokens as language-conditioned prompts. These prompts are temporally fused and inserted into SAM2 to decode pixel-level masks, thereby bridging high-level semantic reasoning and temporally consistent triplet-entity grounding within a unified framework.

$O_{t,n}$, denoted as $M_{t,n}^I$, $M_{t,n}^O \in \{0, 1\}^{H \times W}$. To align with MLLM formulation, both phase labels and triplets are represented as textual categories.

MLLM-Based Reasoning. Given a textual instruction q and video frames $\mathcal{X}$, we first tokenize q into language tokens x_q , and encode $\mathcal{X}$ into visual tokens using a vision encoder followed by a projection to the LLM hidden dimension, producing x_v . The concatenated sequence $[x_q, x_v] \in \mathbb{R}^{L \times D^{\text{llm}}}$ is fed into the LLM as a unified multi-modal input. The LLM auto-regressively generates structured textual outputs x^{llm} . The overall reasoning process is formulated as

$$x^{\text{llm}} = \text{LLM}([\text{Tokenize}(q), \text{VisionEncoder}(\mathcal{X})]). \quad (1)$$

Specifically, for the t -th frame, the generated output x_t^{llm} is structured into two components: a `<think>...</think>` segment that explicitly describes the model’s analysis of the surgical scene, followed by a `<answer>...</answer>` segment that provides the final predictions of the workflow phase and action triplets.

$$x_t^{\text{llm}} = \text{“<think>...</think> \\ <answer> During } \langle p_t \rangle \text{ phase, } \langle N \rangle \text{ surgical triplet(s) is/are identified:} \\ (1) \text{ instrument is } \langle I_{t,n} \rangle \text{ [SEG], target is } \langle O_{t,n} \rangle \text{ [SEG], action is } \langle V_{t,n} \rangle. \\ (2) \dots </answer>”}$$

During fine-tuning, the intrinsic correlations between workflow phases and triplet entities are encoded into the LLM, enabling explicit high-level cross-entity reasoning grounded in the visual context. To further enable spatial grounding, we insert an entity-specific [SEG] token as marker immediately after each triplet-entity token (instrument or target), structured as $\langle I_{t,n} \rangle$ [SEG] and $\langle O_{t,n} \rangle$ [SEG].

These markers function as language-conditioned prompts for downstream segmentation, thereby establishing a direct bridge between high-level semantic reasoning (phase and triplets) and low-level visual grounding.

SAM2-Based Grounding. Given the structured LLM outputs, we first extract all [SEG] tokens and project them into the SAM2 prompt space, yielding [SEG] prompt embeddings $z \in \mathbb{R}^{N_{\text{seg}} \times D_{\text{sam}}}$, where N_{seg} denotes the number of [SEG] markers and D_{sam} is the SAM2 prompt dimension. *Note that for each triplet entity, we assign distinct per-frame [SEG] tokens* rather than using a single shared token across frames. To enforce temporal consistency across per-frame tokens, we group embeddings corresponding to the same entity u (instrument or target) and aggregate them using a residual fusion strategy:

$$\tilde{z}_n = z_n + \text{Proj}\left(\frac{1}{|\mathcal{S}(u)|} \sum_{k \in \mathcal{S}(u)} z_k\right), \quad (2)$$

where $\mathcal{S}(u)$ indexes all occurrences of entity u within the input video. This fusion strategy suppresses per-frame noise, yielding more stable and temporally coherent prompts for SAM2-based mask decoding.

Conditioned on the fused prompt $\tilde{z}$, we first encode the video frames using the SAM2 image encoder and subsequently decode the corresponding segmentation masks. The SAM2 features retain fine-grained spatial details, while the [SEG] prompts carry entity-specific semantic cues derived from the LLM reasoning. By combining these complementary representations, the decoded masks are semantically aligned with the predicted triplet instruments and targets on a frame-by-frame basis, enabling consistent and precise grounding.

Losses. We fine-tune SurgMLLM using the loss function defined as

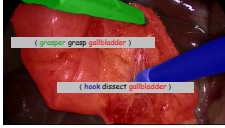
$$\mathcal{L} = \mathcal{L}_{\text{llm}} + \lambda_{\text{bce}} \cdot \mathcal{L}_{\text{bce}} + \lambda_{\text{dice}} \cdot \mathcal{L}_{\text{dice}} + \lambda_{\text{ent}} \cdot \mathcal{L}_{\text{ent}}. \quad (3)$$

The language modeling loss $\mathcal{L}_{\text{llm}}$ is a standard next-token cross-entropy applied to the structured <think> and <answer> outputs. For spatial grounding, we supervise each [SEG]-conditioned mask using a combination of BCE and Dice losses, $\mathcal{L}_{\text{bce}}$ and $\mathcal{L}_{\text{dice}}$, to ensure accurate and robust pixel-level predictions. Additionally, we introduce an entity-aware loss $\mathcal{L}_{\text{ent}}$ that up-weights token positions corresponding to key semantic elements (phase, instrument, verb, and target), encouraging more precise reasoning over critical surgical entities.

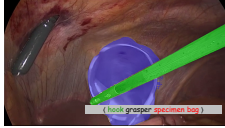
3 Experiments

CholecT45-Scene Dataset. CholecT45 [18] provides frame-level annotations of workflow phase (7 categories) and IVT triplet entities (6 instruments, 10 verbs, 15 targets). *We extend it with pixel-wise mask annotations for both instruments and corresponding targets.* Specifically, representative key-frames within triplet-consistent video segments are annotated using positive/negative point prompts to guide SAM2 mask generation, followed by temporal propagation and manual quality auditing by three expert surgeons with consensus correction.

Dataset	Phase	Triplet	Instrument Mask	Target Mask	Frames
CholecT45 [18]	✓	✓	×	×	90,489
CholecTriplet-Seg [1]	✓	✓	✓	×	30,955
CholecT45-Scene	✓	✓	✓	✓	64,299



«think» The coordination between the hook and the grasper serves to stabilize and ... «/think» «answer» During the **calot triangle dissection** phase, 2 surgical action triplets are identified: (1) the instrument is **grasper**, the target is **gallbladder**, based on the two components, the action is **grasp**; (2) the instrument is **hook**, the target is **gallbladder**, based on the two components, the action is **dissect**. «/answer»



«think» The grasper functions to secure the specimen bag during gallbladder packaging. The immediate physical effect is stabilization of the bag's position relative to the «/think» «answer» During the **gallbladder packaging** phase, 1 surgical action triplet is identified: (1) the instrument is **grasper**, the target is **specimen bag**, based on the two components, the action is **grasp**. «/answer»

Fig. 3. CholecT45-Scene Dataset Overview. Top: Compared with existing surgical datasets, our dataset provides comprehensive annotations including phases, triplets, and triplet-entity masks. Bottom: We visualize annotations in our dataset, presenting instrument and target mask overlays alongside reasoning narratives by Qwen3 [2].

Under deformation, occlusion, or tissue overlap, annotations follow surgical semantic consistency rather than strict exclusive pixel ownership, so independent binary entity masks may overlap. In addition, we generate structured reasoning narratives for each annotated frame, conditioned on the image, masks, and ground-truth phase and triplet labels, enabling supervision for reasoning-based grounding. The resulting CholecT45-Scene contains 64,299 annotated frames from 45 videos, preserving all original phase and triplet labels while adding dense spatial grounding and reasoning annotations for unified surgical scene understanding (see Fig. 3).

Evaluation Metrics. We evaluate three tasks: phase recognition, triplet recognition, and triplet-entity segmentation. For phase recognition, we report video-level Accuracy and phase-level Precision, Recall, and Jaccard under the unrelaxed protocol [13]. Triplet recognition is assessed using official AP-based metrics [19, 10], including component APs (AP_I , AP_V , AP_T), association APs (AP_{IV} , AP_{IT}), and the primary metric AP_{IVT} . For segmentation, we compute IoU for instruments and targets (IoU_I , IoU_T) and report mIoU over all grounded entity masks rather than the arithmetic mean of the two subsets. The instrument/target mask ratio is 0.83/0.17, while IoU_I and IoU_T report the corresponding subsets separately. Following [8], we adopt the official 5-fold cross-validation and report mean $\pm$ standard deviation, with ablations conducted on Fold 1.

Implementation Details. SurgMLLM employs InternVL2.5-4B [5] as MLLM, with both the vision encoder (InternViT) and LLM frozen, and adapts the model using LoRA (rank 128, $\alpha = 256$, dropout 0.05). Input frames are dynamically tiled into 448×448 patches. For grounding, we adopt SAM2-H [22] and fine-tune only the mask decoder on 1024×1024 resized frames. LLM hidden states are projected into the SAM2 prompt space via a two-layer MLP, and temporally fused using a lightweight MLP aggregation module. All trainable parameters are optimized with AdamW with learning rate being 6×10^{-5} . SurgMLLM runs at 2.8 FPS on an NVIDIA RTX PRO 6000.

Quantitative Results. Across all three tasks, SurgMLLM consistently achieves new SOTA performance, demonstrating the effectiveness of unified reasoning and grounding. For *phase recognition* (Tab. 1), it attains the best results on all metrics, notably improving Recall (87.8%) and Jaccard (69.1%), indicating enhanced modeling of workflow transitions. For *triplet recognition* (Tab. 2), SurgMLLM sets a new benchmark on the primary metric AP_{IVT} (46.0% vs. 40.7%), with especially large gains in target AP and association metrics, reflecting stronger *IVT* interaction reasoning. The most significant improvements appear in *triplet-entity segmentation* (Tab. 3), where our method markedly outperforms SAM2-based baselines, achieving 87.8% IoU_I , 67.8% IoU_T , and 84.4% $mIoU$.

These results validate that explicit MLLM-based reasoning combined with [SEG]-guided grounding enables coherent, comprehensive, and spatially consistent surgical scene understanding. In contrast to fragmented pipelines that optimize phase recognition, triplet prediction, and segmentation independently, our unified framework jointly models their intrinsic correlations, leading to more consistent semantic predictions and more reliable dense grounding across tasks.

Table 1. Comparison on phase recognition. Best results are in **bold**.

Method	Video-level	Phase-level		
	Accuracy	Precision	Recall	Jaccard
Trans-SVNet (MICCAI’21) [7]	76.3±2.4	75.1±3.5	74.4±2.4	58.6±2.2
SAHC (TMI’22) [6]	82.0±3.7	78.1±1.9	76.2±3.9	64.0±4.1
SurgFormer (MIA’25) [13]	83.9±2.3	81.1±2.6	77.7±3.9	66.4±4.2
SurgMLLM (Ours)	84.2±0.6	86.3±4.1	87.8±2.6	69.1±3.4

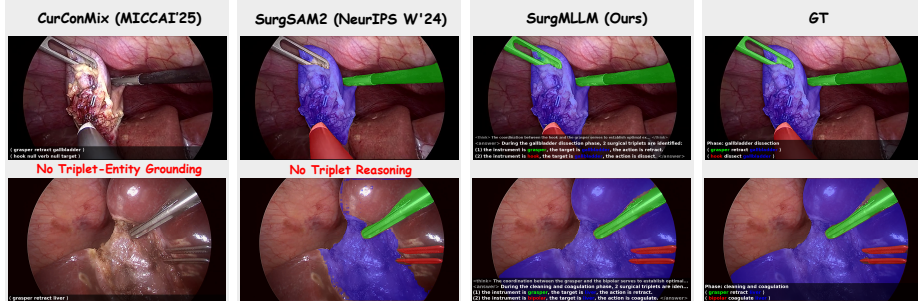
Table 2. Comparison on triplet recognition. Best results are in **bold**.

Method	AP_I	AP_V	AP_T	AP_{IV}	AP_{IT}	AP_{IVT}
RDV (MIA’22) [19]	89.3±2.1	62.0±1.3	40.0±1.4	34.0±3.3	30.8±2.1	29.4±2.8
MT4MTL-KD (TMI’23) [9]	93.1±2.1	71.8±3.4	48.8±3.8	44.9±2.4	43.1±2.0	37.1±0.5
TERL (MICCAI’24) [8]	93.5±2.4	72.8±2.8	51.3±3.8	47.0±5.6	45.7±2.8	38.9±2.5
CurConMix (MICCAI’25) [10]	91.7±2.2	69.5±0.4	51.3±2.9	46.3±5.0	47.1±1.6	40.7±2.1
SurgMLLM (Ours)	89.9±1.9	71.0±6.1	62.2±2.1	68.5±5.8	50.7±2.8	46.0±2.8

Qualitative Results. As shown in Fig. 4, CurConMix [10] struggles to capture fine-grained triplet relations, while SurgSAM2 [15] yields incomplete or inaccurate masks. In contrast, SurgMLLM correctly infers the workflow phase and both triplets, with instrument and target masks closely matching the ground truth, indicating stronger semantic consistency and more reliable dense grounding. More qualitative results are provided in the supplementary video.

Table 3. Comparison on triplet-entity grounding. All SAM-based baselines use a single-point prompt. Best results are in **bold**.

Method	IoU _I	IoU _T	mIoU
SAM2-ZeroShot (ICLR’25) [22]	59.2±0.7	30.5±3.2	47.7±1.2
SAM2-Adapter (ICCV’23) [4]	72.8±1.2	50.3±2.2	63.8±1.0
SurgSAM2 (NeurIPS W’24) [15]	74.2±1.0	53.1±2.6	65.8±1.5
SurgMLLM (Ours)	87.8±1.4	67.8±3.2	84.4±1.2

**Fig. 4. Qualitative comparison.** We compare our SurgMLLM with SOTA methods including CurConMix [10] and SurgSAM2 [15]. Best viewed when zoomed in.

Ablation Study. We conduct ablation studies on Fold 1 of the CholecT45-Scene to examine the effects of joint multi-task learning and the proposed [SEG] prompt design (Tab. 4). Removing supervision for any single task consistently degrades the others, confirming the mutual benefit of phase recognition, triplet reasoning, and grounding under unified optimization. We further evaluate the [SEG] design. Disabling residual fusion (*w/o residual fusion*) or sharing a single token across the clip (*w/o per-frame token*) both reduce performance, demonstrating per-frame semantic tokens combined with temporal fusion are essential for consistent interaction modeling and reliable pixel-level grounding.

Table 4. Ablation study on a subset of tasks within the framework and [SEG] prompt design. Experiments are conducted on Fold 1 of CholecT45-Scene dataset. “-” indicates the corresponding task is not applied.

		Phase Accuracy	Triplet AP _{IVT}	Grounding mIoU
	Ours	84.0	46.4	83.6
Tasks	w/o triplet	80.7	-	80.4
	w/o phase	-	41.8	81.6
	w/o triplet-entity grounding	82.6	40.1	-
[SEG]	w/o residual fusion	82.7	44.2	82.4
	w/o per-frame token	81.5	43.6	81.1

4 Conclusion

This work demonstrates that surgical scene understanding benefits *fundamentally* from unifying temporal modeling, interaction reasoning, and spatial grounding within a single framework rather than treating them as isolated tasks. By coupling structured MLLM-based reasoning with pixel-level grounding, SurgMLLM improves overall surgical scene understanding. The introduced CholecT45-Scene dataset further enables joint evaluation of reasoning and grounding, supporting future research on more integrated surgical video analysis models. Despite these gains, SurgMLLM currently depends on dense triplet-entity mask supervision and may be limited for strict real-time deployment. Future work will explore weakly and semi-supervised grounding and lighter variants.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (No. 62372441), the Guangdong Basic and Applied Basic Research Foundation (No. 2026B1515020093), the Shenzhen Science and Technology Program (Grant Nos. RYX20231211090127030 and JCYJ20250604182948064), and the Research Program of the Department of Education of Guangdong Province (Grant No. 2025ZDZX3027).

Disclosure of Interests. The authors declare no competing interests.

References

1. Alabi, O., Wei, M., Budd, C., Vercauteren, T., Shi, M.: Grounding surgical action triplets with instrument instance segmentation: A dataset and target-aware fusion approach. arXiv preprint arXiv:2511.00643 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Carstens, M., Vasisht, S., Zhang, Z., Barbur, I., Reinke, A., Maier-Hein, L., Hashimoto, D.A., Kolbinger, F.R.: Artificial intelligence for surgical scene understanding: a systematic review and reporting quality meta-analysis. npj Digital Medicine (2025)
4. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: Sam-adapter: Adapting segment anything in underperformed scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
5. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
6. Ding, X., Li, X.: Exploring segment-level semantics for online phase recognition from surgical videos. IEEE Transactions on Medical Imaging (2022)
7. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2021)

8. Gui, S., Wang, Z.: Tail-enhanced representation learning for surgical triplet recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2024)
9. Gui, S., Wang, Z., Chen, J., Zhou, X., Zhang, C., Cao, Y.: Mt4mtl-kd: A multi-teacher knowledge distillation framework for triplet recognition. *IEEE Transactions on Medical Imaging* (2023)
10. Jeon, Y., Shin, J., Park, S., Kim, B., Park, K., Oh, N., Jung, K.H.: Curconmix: A curriculum contrastive learning framework for enhancing surgical action triplet recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2025)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
12. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
13. Lan, M., Si, W., Yan, X., Li, X.: A new dataset and versatile multi-task surgical workflow analysis framework for thoracoscopic mitral valvuloplasty. *Medical Image Analysis* (2025)
14. Li, C., Li, Y., Liu, R., Wang, G., Lv, J., Jin, Y., Si, W., Heng, P.A.: Structural and pixel relation modeling for semisupervised instrument segmentation from surgical videos. *IEEE Transactions on Instrumentation and Measurement* (2023)
15. Liu, H., Zhang, E., Wu, J., Hong, M., Jin, Y.: Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. In: Conference on Neural Information Processing Systems Workshop (2024)
16. Liu, Y., Ma, Z., Pu, J., Qi, Z., Wu, Y., Shan, Y., Chen, C.W.: Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. In: The Thirtieth Annual Conference on Neural Information Processing Systems (2025)
17. Long, Y., Lin, A., Kwok, D.H.C., Zhang, L., Yang, Z., Shi, K., Song, L., Fu, J., Lin, H., Wei, W., et al.: Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. *Science Robotics* (2025)
18. Nwoye, C.I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T., Jia, F., Yang, Y., Wang, H., et al.: Cholectriple2021: A benchmark challenge for surgical action triplet recognition. *Medical Image Analysis* (2023)
19. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis* (2022)
20. Pan, Y., Zou, S.H., Yang, J.W., Si, W.X., Zheng, W.M.: Surgical data science in time-critical contexts: A roadmap toward brain-inspired computing. *Journal of Computer Science and Technology* (2026)
21. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
22. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. In: The Thirteenth International Conference on Learning Representations (2025)
23. Schmidgall, S., Cho, J., Zakka, C., Hiesinger, W.: Gp-vls: A general-purpose vision language model for surgery. *arXiv preprint arXiv:2407.19305* (2024)

24. Wang, G., Bai, L., Wang, J., Yuan, K., Li, Z., Jiang, T., He, X., Wu, J., Chen, Z., Lei, Z., et al.: EndoChat: Grounded multimodal large language model for endoscopic surgery. *Medical Image Analysis* (2025)
25. Yuan, H., Li, X., Zhang, T., Sun, Y., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., et al.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001* (2025)
26. Zeng, Z., Zhuo, Z., Jia, X., Zhang, E., Wu, J., Zhang, J., Wang, Y., Low, C.H., Jiang, J., Zheng, Z., et al.: SurgvLM: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *arXiv preprint arXiv:2506.02555* (2025)