SURGVSI: A BENCHMARK FOR 3D SPATIAL REASONING IN SURGICAL ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust 3D spatial reasoning is essential for understanding complex surgical scenes, where accurate perception of geometric relationships, depth, and motion is critical for downstream decision-making. However, existing benchmarks largely focus on general-domain settings and do not capture the fine-grained geometric interactions and dynamic constraints present in real surgical environments.

In this work, we introduce SurgVSI, a benchmark dataset for evaluating 3D spatial reasoning in surgical environments. SurgVSI is constructed from StereoMIS, EndoVis17, and EndoVis18, and comprises a diverse set of tasks that assess both static and dynamic spatial understanding, including depth estimation, distance estimation, camera ego-motion analysis, and instrument path tracking and completion. The dataset contains 5,998 evaluation instances and 147,386 training instances, with metric annotations derived from stereo reconstruction and camera calibration.

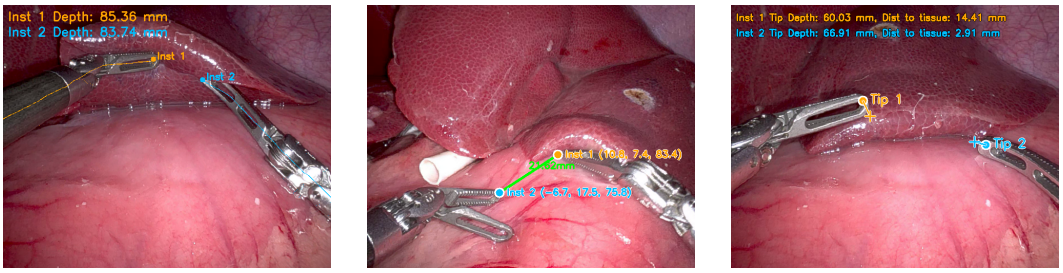
We conduct extensive evaluations across representative models to characterize current capabilities on SurgVSI, revealing substantial challenges in accurate geometric reasoning and dynamic spatial understanding. Our benchmark establishes a standardized testbed for studying 3D spatial reasoning in surgical contexts and provides a foundation for future research in perception, reasoning, and robotic systems operating in complex real-world environments.

1 INTRODUCTION

Autonomous robotic surgery and high-precision intraoperative assistance represent transformative advancements in modern healthcare, enabling minimally invasive procedures with enhanced accuracy and reduced human error (Shademan et al., 2016; Twinanda et al., 2016; Allan et al., 2020). At the core of these systems lies the need for sophisticated *3D spatial reasoning*—the ability to perceive and reason about geometric relationships, absolute and relative depths, camera ego-motion, and dynamic trajectories of surgical instruments relative to anatomical structures in real time. Such capabilities are essential not only for safe autonomous manipulation but also for interpretable decision support, surgical skill assessment, and next-generation navigation systems operating under the stringent constraints of the operating room.

Recent years have witnessed remarkable progress in general-domain vision-language models (VLMs) and reasoning techniques. Chain-of-Thought prompting (Wei et al., 2022) and its variants—including Self-Consistency (Wang et al., 2022), Least-to-Most (Zhou et al., 2023), and complexity-based prompting (Fu et al., 2023)—have significantly improved multi-step reasoning in large language models. Zero-shot reasoning capabilities have been further enhanced (Kojima et al., 2022; Zhang et al., 2022), while more structured paradigms such as Tree-of-Thoughts (Yao et al., 2023), Graph-of-Thoughts (Besta et al., 2023), and Graph-based agent planning (Wu et al., 2025a) enable deliberate problem solving. Agentic frameworks that synergize reasoning and acting (Yao et al., 2022; Shinn et al., 2023), tool-augmented models (Schick et al., 2023; Qin et al., 2023; Patil et al., 2023), and reinforcement learning approaches for reasoning (Lightman et al., 2023; Zhang et al., 2025b; Yue et al., 2025; Zhang et al., 2025a; Xu et al., 2025) have pushed the frontier of complex inference. Additional advances include automatic prompt augmentation (Shum et al., 2023), amortized inference (Hu et al., 2024), and multi-agent collaboration (Wang et al., 2024; Wu et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Instrument absolute depth estimation

(b) Inter-instrument distance estimation

(c) Instrument-to-tissue distance estimation

(d) Camera ego-motion analysis

(e) Instrument path tracking

Figure 1: Representative examples from the SurgVSI benchmark illustrating its core 3D spatial reasoning tasks. The top row shows static spatial reasoning: (left) instrument absolute depth estimation, (middle) inter-instrument distance estimation, and (right) instrument-to-tissue distance estimation. The bottom row demonstrates dynamic spatial reasoning: (left) camera ego-motion analysis and (right) instrument path tracking under tissue deformation. These examples highlight the geometric complexity of surgical scenes.

Despite these advances, current VLMs face substantial challenges in quantitative spatial understanding when applied to surgical scenes. The challenges are multifaceted: surgical environments are highly dynamic with frequent occlusions, specular reflections, and rapid camera motion; instruments and tissues exhibit complex 3D interactions; and accurate depth and motion estimation must be robust to varying lighting, tissue deformation, and partial visibility. Existing benchmarks for spatial reasoning (Ma et al., 2025; Yang et al., 2024; Fu et al., 2025) predominantly target everyday or synthetic scenes and do not capture the fine-grained geometric interactions and clinical constraints required in surgery. Consequently, there remains a critical gap in standardized evaluation protocols for 3D spatial intelligence in surgical contexts.

In this work, we introduce **SurgVSI**, a comprehensive benchmark dataset specifically designed to evaluate 3D spatial reasoning in real surgical environments. SurgVSI is constructed from stereo endoscopic videos and images, with metric annotations derived from stereo reconstruction, camera calibration, and task-specific geometric computation. The benchmark comprises six core tasks that systematically probe both static and dynamic aspects of spatial understanding:

- **Instrument absolute depth estimation**: Predicting the optical-axis depth of surgical instrument tips.
- **Inter-instrument distance estimation**: Estimating 3D Euclidean distances between multiple instrument tips.
- **Instrument-to-tissue distance estimation**: Measuring the minimum 3D distance from instrument tips to valid local background geometry.
- **Camera ego-motion analysis**: Quantifying translation and rotation of the endoscope between frames.
- **Instrument path tracking**: Reconstructing 3D trajectories of instrument tips over short temporal windows.
- **Instrument path completion**: Estimating intermediate 3D displacements given endpoint images and coordinates.

These tasks are instantiated as VLM-compatible question-answer pairs (image or image-sequence inputs paired with precise numeric or structured outputs) with carefully designed prompts that emphasize geometric reasoning. The benchmark covers three surgical data sources and six tasks, with separate training and evaluation manifests. Known query coordinates align the target across models; the tasks evaluate geometry conditioned on localization, rather than unrestricted instrument detection. Here, spatial reasoning includes geometric perception, numerical estimation, and temporal inference.

We conduct evaluations of representative general-purpose and spatial VLMs, together with geometry-assisted surgical baselines on SurgVSI, revealing substantial performance gaps in accurate geometric perception and dynamic spatial understanding. These results highlight the unique challenges of surgical 3D reasoning and establish SurgVSI as a rigorous, standardized testbed for future research.

SurgVSI and its geometric question-answer interface aim to accelerate progress toward robust perception and reasoning systems for robotic surgery. SurgVSI provides a foundation for developing more reliable VLMs, advancing surgical AI safety, and ultimately enabling the next generation of intelligent intraoperative assistance.

2 Related Work

2.1 Visual Spatial Reasoning Benchmarks

Spatial reasoning has been extensively studied in general-domain computer vision and vision-language models. Early benchmarks such as Visual Question Answering (VQA) (Antol et al., 2015) and visual grounding tasks primarily focus on 2D spatial relations. More recent efforts have advanced toward 3D understanding, including 3D-QA datasets, spatial reasoning benchmarks on synthetic scenes (Ma et al., 2025), and embodied navigation tasks that require camera motion and trajectory reasoning (Anderson et al., 2017; Savva et al., 2019).

However, these benchmarks are largely built on everyday objects or simulated environments that do not reproduce the full combination of surgical imaging conditions. They do not capture the unique challenges of surgical scenes: highly dynamic motion, specular reflections, partial occlusions by instruments, tissue deformation, and clinical constraints. Dedicated surgical evaluation is therefore needed to characterize how model capabilities transfer to these spatial tasks.

2.2 Surgical Vision and Robotics Datasets

Several datasets have been proposed for surgical scene understanding, including instrument segmentation (Allan et al., 2020), phase recognition and action recognition in endoscopic videos (Twinanda et al., 2016), as well as skill assessment and multi-modal robotic surgery simulation. Stereo endoscopic datasets have also been released to support depth estimation and 3D reconstruction (Hayoz et al., 2023).

Nevertheless, existing surgical datasets primarily target *2D perception* (instrument segmentation (Allan et al., 2020), detection, phase recognition and action recognition (Twinanda et al., 2016)) or low-level 3D reconstruction (Hayoz et al., 2023). These resources motivate a complementary VLM-compatible benchmark for numerical 3D spatial reasoning tasks such as instrument absolute depth estimation, inter-instrument distance estimation, instrument-to-tissue distance estimation, camera ego-motion analysis, or instrument path completion. SurgVSI fills this critical gap by offering a VLM-compatible benchmark with metric annotations derived from real stereo endoscopic videos.

2.3 Reasoning in Vision-Language Models

The rise of large language models has spurred significant advances in reasoning techniques. Chain-of-Thought (CoT) prompting (Wei et al., 2022) and its variants—including Self-Consistency (Wang et al., 2022), Least-to-Most (Zhou et al., 2023), complexity-based prompting (Fu et al., 2023), automatic CoT generation (Zhang et al., 2022; Shum et al., 2023), and Program-of-Thoughts (Chen et al., 2022)—have greatly improved multi-step inference. Zero-shot reasoning capabilities have

Table 1: Statistics of SurgVSI per task. Counts are question-answer instances.

Task	Training	Evaluation
Instrument absolute depth estimation	25,055	1,000
Camera ego-motion analysis	50,673	1,000
Inter-instrument distance estimation	6,781	998
Instrument-to-tissue distance estimation	24,777	1,000
Instrument path tracking	20,050	1,000
Instrument path completion	20,050	1,000
Total	147,386	5,998

been further enhanced (Kojima et al., 2022). Structured reasoning paradigms such as Tree-of-Thoughts (Yao et al., 2023; Long, 2023), Graph-of-Thoughts (Besta et al., 2023), and topology-aware reasoning (Minegishi et al., 2025) enable more deliberate problem solving. Agentic frameworks that synergize reasoning and acting (Yao et al., 2022; Shinn et al., 2023), tool-augmented models (Schick et al., 2023; Qin et al., 2023; Patil et al., 2023), and reinforcement learning approaches for reasoning (Lightman et al., 2023; Zhang et al., 2025b; Yue et al., 2025; Zhang et al., 2025a; Xu et al., 2025) continue to push the frontier of complex inference.

Additional advances include multi-agent collaboration (Wang et al., 2024; Wu et al., 2024), amortized inference (Hu et al., 2024), instruction tuning (Ouyang et al., 2022; Sanh et al., 2022), and various agent benchmarks and environments (Liu et al., 2025; Zhou et al., 2024).

Despite these advances, most reasoning methods remain primarily linguistic or 2D-image grounded. Their progress does not by itself establish accurate surgical 3D spatial understanding, motivating benchmarks that evaluate numerical geometry explicitly. SurgVSI addresses this by providing a rigorous testbed that requires explicit geometric modeling and precise spatial inference.

With SurgVSI, we aim to bridge the gap between general reasoning progress and the demanding requirements of surgical spatial intelligence, fostering the development of more reliable and clinically applicable vision-language systems.

3 Benchmark Construction

We introduce **SurgVSI**, a large-scale benchmark dataset specifically designed to evaluate 3D spatial reasoning in real surgical environments. The dataset is constructed from real stereo endoscopic videos through an automated geometric processing pipeline. This section details the data source and the multi-stage construction process.

3.1 Data Source

SurgVSI is built upon StereoMIS, EndoVis17, and EndoVis18 (Hayoz et al., 2023; Allan et al., 2019; 2020), which contain stereo endoscopic data acquired with da Vinci surgical systems during porcine procedures. StereoMIS provides stereo-video sequences captured with the da Vinci Xi platform; EndoVis17 and EndoVis18 provide stereo images, calibration, and instrument or surgical-scene annotations. The data exhibit realistic surgical dynamics, including tissue deformation, instrument motion, and partial occlusions. Table 6 in Appendix A summarizes acquisition, source scale, annotations, and the contribution of each dataset.

The evaluation set uses StereoMIS P2_7, P2_8, and P3; EndoVis18 sequences 2, 5, 9, and 15; and ten EndoVis17 test identifiers. EndoVis17 follows the official split: frames 000–224 of recordings 1–8 are used for training, while their remaining 75 frames and all 300 frames of recordings 9–10 are used for testing. Full sequence and task inventories are given in Appendix A.

3.2 Data Processing Pipeline

The SurgVSI benchmark is generated through a systematic multi-stage pipeline that transforms raw stereo videos into metric 3D geometric annotations.

1. Stereo Frame Extraction and Alignment. Synchronized left and right frames are extracted from the original video streams while maintaining exact temporal correspondence with the provided instrument segmentation masks. The working image dimensions and calibration intrinsics are kept consistent for subsequent geometric processing.

2. Metric Depth Map Computation. Dense depth maps are computed from the rectified stereo image pairs using a stereo matching algorithm based on semi-global block matching. Edge-aware disparity refinement is applied to improve depth completeness in challenging regions such as specular highlights and low-texture tissue surfaces. The resulting depth values are reprojected into 3D space using the sequence-specific stereo calibration parameters, producing metric-scale depth maps aligned with the left camera coordinate system. Invalid disparity estimates are excluded from valid-depth computations.

3. Temporal Depth Refinement. To enhance temporal consistency and suppress noise inherent in per-frame stereo matching, a sliding-window temporal pooling is applied across consecutive frames. For each pixel location, depth values from neighboring frames within a small temporal window are aggregated using a median filter. This operation suppresses temporal depth fluctuations in the reconstructed maps.

4. Instrument Tip Localization and 3D Reconstruction. Instrument tips are localized in each frame by first applying connected-component analysis to the binary segmentation masks to isolate individual instruments (discarding small noise regions with area below a threshold). For each instrument region, morphological skeletonization via thinning is performed, followed by endpoint detection along the tool shaft. The tip location is determined as the skeleton endpoint closest to the image center using a combination of morphological heuristics and local depth consistency checks. These 2D tip coordinates are then back-projected into 3D space using the refined depth map and scaled camera intrinsics. Specifically, given a 2D tip pixel (u, v) with corresponding depth value Z, the 3D coordinates are recovered as:

$$X = (u - c_x) \cdot \frac{Z}{f_x}, \quad Y = (v - c_y) \cdot \frac{Z}{f_y},$$

where (f_x, f_y) and (c_x, c_y) are the focal lengths and principal point of the left camera, respectively (scaled to the depth map resolution). Both resized and original-resolution 2D coordinates are retained for downstream tasks.

5. Instrument-to-tissue distance estimation. For each detected instrument tip, the minimum 3D distance to the nearest background tissue is computed. The background mask is obtained by inverting the instrument mask and applying morphological dilation. All valid background pixels (non-instrument regions with reliable depth) within a local cylindrical neighborhood around the tip—defined by a spatial radius constraint and a depth-band constraint—are considered. For each candidate background point, its 3D coordinate is reconstructed using the same back-projection formula. The distance is defined as the minimum Euclidean distance from the instrument tip to these background points. It measures separation from the locally reconstructed background. The QA prompt specifies both the instrument pixel and the selected tissue pixel.

6. Camera Ego-Motion and Instrument Path Construction. Camera ego-motion between selected frame pairs is estimated by extracting robust SIFT features from background regions (instrument pixels are masked out). Feature matching is performed using FLANN with Lowe's ratio test. 3D points from the reference frame are reconstructed via depth back-projection, and relative camera pose is solved using the perspective-n-point algorithm with RANSAC. This yields estimated translation vectors (in physical units) and rotation (converted to Euler angles). Instrument trajectories are constructed by linking the reconstructed 3D tip positions across temporally consistent frame windows, producing short relative 3D motion paths. Each path is expressed relative to its initial tip position in camera coordinates.

3.3 Task Construction and Dataset Statistics

From the geometric primitives obtained above (3D tip positions, distances, camera motions, and trajectories), we automatically generate VLM-compatible question-answer pairs for the following

six core spatial reasoning tasks: instrument absolute depth estimation, inter-instrument distance estimation, instrument-to-tissue distance estimation, camera ego-motion analysis, instrument path tracking, and instrument path completion.

Each task instance consists of one or more input images (or short image sequences) paired with a carefully designed prompt that provides geometric context, followed by a numeric or structured reference annotation. The benchmark contains 5,998 evaluation instances and a training pool of 147,386 instances (Table 1). Five tasks contain 1,000 evaluation instances each: 300 from EndoVis18, 300 from EndoVis17, and 400 from StereoMIS. Inter-instrument distance requires at least two visible instruments and contains 998 instances, distributed as 41, 217, and 740 across these sources.

The resulting SurgVSI benchmark enables rigorous, reproducible evaluation of 3D spatial reasoning capabilities under realistic surgical conditions, providing a standardized testbed for vision-language models in clinically relevant scenarios. Tracking provides seven observed images and requests seven relative 3D points. Completion provides the first and last images with start/end relative coordinates, and requests seven points including the endpoints. Prompts specify geometric context, target coordinates, and output schemas.

The sources broaden instrument and scene coverage; source-label, visibility, and confidence distributions are provided in Appendix C. Answer–metadata, coordinate-projection, and trajectory-structure checks assess the internal consistency of these annotations.

3.4 Evaluation Metrics

Evaluation uses calibrated stereo-derived geometric references from the three data sources. Distance-related metrics are in millimeters and rotation quantities in degrees. Full formulas and aggregation rules appear in Appendix B.

Metrics of distance estimation For instrument depth, inter-instrument distance, and tissue distance, we report mean absolute error (MAE), root mean squared error (RMSE), the population standard deviation of absolute errors (σ_1), and the population standard deviation of absolute relative errors (σ_2, in percent). Zero references are excluded only from relative-error statistics.

Metrics of motion analysis For camera motion, T-RMS and T-MAE are the RMSE and MAE of translation-vector magnitude differences; E-RMS is the RMSE of Euler-vector magnitude differences. Metrics are computed directly in the camera coordinate system.

Metrics of tracking and completion Tracking uses mean per-tip position error (MPTE), PCK@2 mm, and Group-RMS. Group-RMS averages, across instrument-label groups, the RMS of trajectory-mean point errors. Completion uses mean path deviation (MeanDev) and mean absolute path-length deviation (LenDev). Only complete seven-point paths are accepted; point-error metrics use the six noninitial points and LenDev uses the full path.

4 Experiment

4.1 Experimental Results and Analysis

We evaluate five general-purpose VLMs (GPT-5.5, Qwen3.7-Plus, Kimi-K2.6, Claude-Opus-4.8, and Gemini-3.5-Flash), the spatial VLMs ViLaSR (Wu et al., 2025b) and VST-7B-RL (Yang et al., 2026), and two surgical methods, Endo3R (Guo et al., 2026) and Surgical SAM 2 (Liu et al., 2024). VLMs receive the image inputs and task queries through their respective adapters. Results are summarized in Tables 2–4; Table 5 reports the geometry-assisted surgical baselines separately.

All seven VLMs are evaluated using the same geometric references and task-specific metrics. Source-specific results and scoring counts are provided in Appendix D.

Table 2: Quantitative results of representative VLMs on SurgVSI for instrument absolute depth estimation and camera ego-motion analysis. Camera columns measure translation/Euler-vector magnitude errors. Relative-error dispersion is percent. Bold marks the best value per column; valid-output counts appear in Appendix D.

Model	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓	T-RMS ↓	E-RMS ↓	T-MAE ↓
GPT-5.5	36.89	63.87	52.14	127.31	6.31	5.59	1.89
Qwen3.7-Plus	28.75	57.61	49.92	92.36	8.57	5.69	3.48
Kimi-K2.6	52.67	114.46	101.62	184.53	7.09	5.82	2.57
Claude-Opus-4.8	55.60	100.71	83.97	183.73	11.35	5.62	9.30
Gemini-3.5-Flash	**28.27**	**53.90**	**45.89**	88.48	6.57	5.81	2.09
ViLaSR	39.14	78.00	67.47	51.29	83.17	7.22	39.96
VST-7B-RL	61.49	80.33	51.69	**2.06**	**2.85**	**2.98**	**1.50**

4.2 Static Spatial Reasoning Tasks

- **Instrument absolute depth estimation** (Table 2, left): Errors remain substantial across the evaluated VLMs, with MAE ranging from 28.27 mm to 61.49 mm. RMSE ranges from 53.90 mm to 114.46 mm, indicating a substantial spread of depth errors.
- **Inter-instrument distance estimation** (Table 3, left): MAE ranges from 30.36 mm to 54.89 mm across models. The differences between MAE and RMSE indicate that large individual errors contribute strongly to aggregate performance.
- **Instrument-to-tissue distance estimation** (Table 3, right): MAE is smaller in absolute units, ranging from 0.88 mm to 7.25 mm, but relative-error dispersion remains high. Target scale and visibility differ across distance tasks, and near-zero references amplify relative errors.

Table 3: Quantitative results on inter-instrument distance estimation (left) and instrument-to-tissue distance estimation (right). Absolute errors are mm; relative-error dispersion is percent. Bold marks the best value per column.

Model	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
GPT-5.5	49.73	93.70	79.42	162.55	3.71	28.59	28.34	65.00
Qwen3.7-Plus	54.14	94.05	76.91	131.10	4.14	28.54	28.24	371.77
Kimi-K2.6	37.90	89.96	81.58	143.44	4.24	28.61	28.30	579.68
Claude-Opus-4.8	38.79	86.59	77.42	133.19	7.25	29.05	28.13	5228.20
Gemini-3.5-Flash	54.89	94.34	76.73	221.76	4.59	28.45	28.08	889.39
ViLaSR	**30.36**	**67.81**	**60.63**	185.59	4.98	30.29	29.88	965.49
VST-7B-RL	39.41	93.36	84.64	**3.09**	**0.88**	**1.63**	**1.37**	**17.87**

4.3 Dynamic Spatial Reasoning Tasks

- **Camera ego-motion analysis** (Table 2, right): For the five general-purpose VLMs, translation-magnitude RMSE ranges from 6.31 to 11.35 mm and Euler-magnitude RMSE from 5.59 to 5.82 degrees. ViLaSR and VST have translation-magnitude RMSE of 83.17 mm and 2.85 mm, respectively.
- **Instrument path tracking** (Table 4, left): Performance varies substantially. GPT-5.5, Claude-Opus-4.8, and Gemini-3.5-Flash have MPTE of 31.25, 34.50, and 32.99 mm, respectively, while the range across all seven VLMs extends to 520.12 mm. PCK@2 mm remains between 0.79% and 14.62%, indicating limited fine-scale agreement with reference paths.
- **Instrument path completion** (Table 4, right): MeanDev for the five general-purpose VLMs lies between 19.40 and 19.52 mm, with LenDev approximately 103 mm. ViLaSR and VST have MeanDev of 47.36 mm and 30.63 mm, respectively. These results high-

light the challenge of reconstructing intermediate instrument motion from the supplied endpoints.

Table 4: Quantitative results on instrument path tracking (left) and instrument path completion (right). PCK is percent; other values are mm. Group-RMS uses instrument-label groups. Bold marks the best value per column.

Model	MPTE ↓	PCK ↑	Group-RMS ↓	MeanDev ↓	LenDev ↓
GPT-5.5	**31.25**	13.59	83.73	**19.40**	102.66
Qwen3.7-Plus	194.38	10.92	475.59	19.46	102.78
Kimi-K2.6	411.91	5.28	672.48	19.49	**102.58**
Claude-Opus-4.8	34.50	2.69	83.50	19.41	102.76
Gemini-3.5-Flash	32.99	**14.62**	**80.58**	19.52	103.09
ViLaSR	520.12	0.79	608.04	47.36	120.53
VST-7B-RL	188.95	7.03	262.80	30.63	129.68

Geometry-assisted specialist baselines. Endo3R reports depth MAE of 14.19 mm and inter-instrument distance MAE of 24.60 mm. Surgical SAM 2 reports tracking MPTE of 38.22 mm, PCK@2 mm of 20.69%, and Group-RMS of 108.22 mm (Table 5). Endo3R's adapter scales predictions using reference-depth medians. Surgical SAM 2 combines mask propagation with stereo depth, calibration, and geometric tip extraction. Appendix E describes the adapter settings and shared tracking metrics.

Table 5: Results of geometry-assisted specialist baselines. These methods use auxiliary depth/calibration inputs (Appendix E), limiting direct comparison with Tables 2–4. Absolute quantities are mm, rotation is degrees, and σ_2/PCK are percent. SAM 2 uses the same tracking metrics as Table 4, on 932 complete trajectories (5,592 noninitial points).

Endo3R task	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
Depth	14.19	50.33	48.29	31.56
Inter-instrument	24.60	87.35	83.82	28.12
Tissue distance	5.64	53.50	53.20	316.04
Endo3R task	**RPE ↓**	**Rot.RPE ↓**	**Abs.Err ↓**	
Camera	9.47	8.83	6.36	
Surgical SAM 2 task	**MPTE ↓**	**PCK@2mm ↑**	**Group-RMS ↓**	
Tracking	38.22	20.69	108.22	

4.4 Limitations

While SurgVSI provides a valuable benchmark for 3D spatial reasoning in surgical environments, it has several limitations. The dataset is derived solely from porcine procedures using da Vinci systems, which may not fully capture the variability of human anatomy and pathology. Geometric annotations inherit residual errors from stereo reconstruction in challenging cases such as specular reflections, occlusions, and fast motion. Our validation assesses internal consistency; independent reference measurements are needed to quantify physical accuracy. Supervised fine-tuning and surgical-specific pretraining remain directions for future work, along with expanding procedural diversity and scale.

5 Conclusion

In this work, we introduce **SurgVSI**, a comprehensive benchmark for 3D spatial reasoning in real surgical environments. Built from stereo endoscopic data across StereoMIS, EndoVis17, and En-

doVis18, SurgVSI provides metric geometric annotations for six core tasks covering both static and dynamic spatial understanding.

Extensive evaluations of general-purpose and spatial vision-language models reveal significant performance gaps, particularly in fine-grained depth estimation and dynamic instrument tracking. Geometry-assisted surgical baselines further complement the evaluation of these capabilities.

SurgVSI provides a standardized testbed to drive progress in surgical AI. Future work includes supervised adaptation, multi-frame reasoning, surgical-domain pretraining, and hybrid geometric approaches for intelligent intraoperative assistance.

Ethics Statement

This work reuses public porcine surgical datasets and does not report a new patient study. Access to and redistribution of source imagery are governed by the respective dataset terms. SurgVSI is intended for research evaluation and does not establish deployment or clinical safety.

Reproducibility Statement

The supplementary material includes the SurgVSI dataset annotations for all 5,998 evaluation and 147,386 training instances, together with the data-construction and evaluation code. The package provides task questions, reference answers, geometric metadata, image paths, the answer parser, and metric implementations, with usage instructions in the README. The appendices describe the data sources, evaluation protocol, and implementation details.

AI Use Statement

AI assistance was used for language editing, LaTeX preparation, code inspection, data-inventory verification, and analysis of model outputs. The authors are responsible for the methods, analyses, and conclusions.

References

M. Allan, Alexey A. Shvets, T. Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, S. Bodenstedt, et al. 2017 robotic instrument segmentation challenge, 2019. URL https://arxiv.org/abs/1902.06426.

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge, 2020. URL https://arxiv.org/abs/2001.11190.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, I. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683. IEEE, 2017. doi: 10.1109/CVPR.2018.00387.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Journal of Computer Vision*, volume 123, pp. 4–31. Springer Science and Business Media LLC, 2015. doi: 10.1007/s11263-016-0966-6.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models, 2023. URL https://arxiv.org/abs/2308.09687.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Trans. Mach. Learn. Res.*, 2022.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive, 2025. URL https://arxiv.org/abs/2404.12390.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning, 2023. URL https://arxiv.org/abs/2210.00720.

Jiaxin Guo, Wenzhen Dong, Tianyu Huang, Hao Ding, Ziyi Wang, Haomin Kuang, Qi Dou, and Yun-Hui Liu. Endo3R: Unified online reconstruction from dynamic monocular endoscopic video, 2026. URL https://arxiv.org/abs/2504.03198.

Michel Hayoz, Christopher Hahne, Mathias Gallardo, Daniel Candinas, Thomas Kurmann, Maximilian Allan, and Raphael Sznitman. Learning how to robustly estimate camera pose in endoscopic videos. *International Journal of Computer Assisted Radiology and Surgery*, 18(7):1185–1192, 2023. doi: 10.1007/s11548-023-02919-w. URL https://link.springer.com/article/10.1007/s11548-023-02919-w.

Edward J. Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Y. Bengio, and Esmeralda S. Whitammer. Amortizing intractable inference in large language models, 2024. URL https://arxiv.org/abs/2310.04363.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2022. URL https://arxiv.org/abs/2205.11916.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Haofeng Liu, Erli Zhang, Junde Wu, Mingxuan Hong, and Yueming Jin. Surgical SAM 2: Real-time segment anything in surgical video by efficient frame pruning, 2024. URL https://arxiv.org/abs/2408.07931.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents, 2025. URL https://arxiv.org/abs/2308.03688.

Jieyi Long. Large language model guided tree-of-thought, 2023. URL https://arxiv.org/abs/2305.08291.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jie-Neng Chen, Celso de Melo, and Alan L. Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934. IEEE, 2025. doi: 10.1109/iccv51701.2025.00651.

Gouki Minegishi, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Topology of reasoning: Understanding large reasoning models through reasoning graph properties, 2025. URL https://arxiv.org/abs/2506.05744.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, S. Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Neural Information Processing Systems*, 35:27730–27744, 2022. doi: 10.52202/068431-2011.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive apis, 2023. URL https://arxiv.org/abs/2305.15334.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis, 2023. URL https://arxiv.org/abs/2307.16789.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization, 2022. URL https://arxiv.org/abs/2110.08207.

M. Savva, Jitendra Malik, Devi Parikh, Dhruv Batra, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, et al. Habitat: A platform for embodied ai research. In *IEEE International Conference on Computer Vision*, pp. 9338–9346. IEEE, 2019. doi: 10.1109/ICCV.2019.00943.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, R. Raileanu, M. Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023. URL https://arxiv.org/abs/2302.04761.

Azad Shademan, Ryan S Decker, Justin D Opfermann, Simon Leonard, Axel Krieger, and Peter CW Kim. Supervised autonomous robotic soft tissue surgery. *Science translational medicine*, 8(337): 337ra64–337ra64, 2016. doi: 10.1126/scitranslmed.aad9398.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

KaShun Shum, Shizhe Diao, and Tong Zhang. Automatic prompt augmentation and selection with chain-of-thought from labeled data, 2023. URL https://arxiv.org/abs/2302.12822.

A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. Mathelin, and N. Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016. doi: 10.1109/TMI.2016.2593957.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities, 2024. URL https://arxiv.org/abs/2406.04692.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2022. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022. URL https://arxiv.org/abs/2201.11903.

Jiaqi Wu, Qinlao Zhao, Zefeng Chen, Kai Qin, Yifei Zhao, Xueqian Wang, and Yuhang Yao. Gap: Graph-based agent planning with parallel tool use and reinforcement learning. *arXiv.org*, 2025a. doi: 10.48550/arXiv.2510.25320.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing, 2025b. URL https://arxiv.org/abs/2506.09965.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *arXiv.org*, 2024.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. Towards large reasoning models: A survey of reinforced reasoning with large language models, 2025. URL https://arxiv.org/abs/2501.09686.

Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Computer Vision and Pattern Recognition*, pp. 10632–10643. IEEE, 2024. doi: 10.1109/CVPR52734.2025.00994. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Thinking_in_Space_How_Multimodal_Large_Language_Models_See_Remember_CVPR_2025_paper.html.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning, 2026. URL https://arxiv.org/abs/2511.05491.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2022. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. URL https://arxiv.org/abs/2305.10601.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models, 2025a. URL https://arxiv.org/abs/2509.08827.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning, 2025b. URL https://arxiv.org/abs/2501.07301.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models, 2022. URL https://arxiv.org/abs/2210.03493.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models, 2023. URL https://arxiv.org/abs/2205.10625.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents, 2024. URL https://arxiv.org/abs/2307.13854.

A Data Inventory

Table 6: Source datasets and their contribution to SurgVSI. The upper panel describes the public sources; the lower panel reports our derived inventory.

	StereoMIS	**EndoVis17**	**EndoVis18**
Source reference	Hayoz et al. (2023)	Allan et al. (2019)	Allan et al. (2020)
Acquisition	In-vivo surgery; 3 porcine subjects	Porcine robot-assisted nephrectomy	Porcine training procedures
Surgical platform	da Vinci Xi	da Vinci Xi	da Vinci X or Xi
Public source scale	11 stereo-video sequences	10 recordings; 3,000 stereo pairs	19 sequences; 5,700 stereo pairs
Source train/test allocation	1 / 10 sequences	1,800 / 1,200 stereo pairs	15 / 4 sequences; 4,500 / 1,200 pairs
Source information	Stereo video, calibration, camera poses	Stereo images, calibration, instrument masks	Stereo images, calibration, instrument and anatomy labels
SurgVSI derived inventory (QA totals)			
Sequence IDs (train / eval.)	8 / 3	8 / 10[a]	11 / 4
Training QA	132,779	7,984	6,623
Evaluation QA	2,740	1,717	1,541
Total QA	135,519	9,705	8,164

StereoMIS scope and source allocation refer to its public porcine release; its private human data are not used. EndoVis pair counts are calculated from the source publications, and manual labels concern the left-eye images. [a]EndoVis17 follows the official frame-based train/test split; its reported tr/te IDs are derived identifiers, not counts of independent recordings. The StereoMIS and EndoVis18 SurgVSI partitions are distinct from the source allocations listed above. QA counts are not unique-frame or procedure counts. Sequence-level accounting is documented in Appendix A.

Counts refer to generated QA instances. Direct enumeration of the benchmark files verifies 5,998 evaluation instances, comprising 1,541 from EndoVis18, 1,717 from EndoVis17, and 2,740 from StereoMIS. The training files contain 147,386 instances. The task-level and per-sequence inventories below use these verified counts.

Table 7: Training instances by task and source, from the task inventory.

Task	StereoMIS	EndoVis18	EndoVis17	Total
Instrument absolute depth	21,211	1,719	2,125	25,055
Inter-instrument distance	6,076	192	513	6,781
Instrument-to-tissue distance	20,940	1,717	2,120	24,777
Camera ego-motion	47,450	1,549	1,674	50,673
Instrument path tracking	18,551	723	776	20,050
Instrument path completion	18,551	723	776	20,050
Total	132,779	6,623	7,984	147,386

Table 8: Evaluation instances by task and source, from the task inventory. The inter-instrument task uses availability-based sampling.

Task	EndoVis18	EndoVis17	StereoMIS	Total
Instrument absolute depth	300	300	400	1,000
Inter-instrument distance	41	217	740	998
Instrument-to-tissue distance	300	300	400	1,000
Camera ego-motion	300	300	400	1,000
Instrument path tracking	300	300	400	1,000
Instrument path completion	300	300	400	1,000
Total	1,541	1,717	2,740	5,998

A.1 SEQUENCE INVENTORY AND SOURCE MAPPING

The sequence identifiers and counts correspond to the final manifests. EndoVis17 te2 corresponds to test_instrument_dataset_2 and contains 153 evaluation instances. StereoMIS P2_3 contributes 4,205 instances to the training total of 132,779. EndoVis17 training sequences tr1–tr8 use source frames 000–224; test sequences te1–te8 use frames 225–299, and te9–te10 use frames 000–299 (Allan et al., 2019).

Table 9: Training and evaluation sequence inventories. Evaluation rows sum to 5,998, matching the task-level total. EndoVis17 te2 contains 153 instances.

Training

Dataset	Sequence	Instances
StereoMIS	P1	21,453
StereoMIS	P2_0	41,589
StereoMIS	P2_1	4,428
StereoMIS	P2_2	4,839
StereoMIS	P2_3	4,205
StereoMIS	P2_4	5,256
StereoMIS	P2_5	6,014
StereoMIS	P2_6	44,995
EndoVis18	seq1	934
EndoVis18	seq3	993
EndoVis18	seq4	535
EndoVis18	seq6	527
EndoVis18	seq7	480
EndoVis18	seq10	557
EndoVis18	seq11	462
EndoVis18	seq12	517
EndoVis18	seq13	474
EndoVis18	seq14	531
EndoVis18	seq16	613
EndoVis17	tr1	899
EndoVis17	tr2	995
EndoVis17	tr3	975
EndoVis17	tr4	1,140
EndoVis17	tr5	739
EndoVis17	tr6	1,337
EndoVis17	tr7	686
EndoVis17	tr8	1,213

Evaluation

Dataset	Sequence	Instances
StereoMIS	P2_7	2,250
StereoMIS	P2_8	389
StereoMIS	P3	101
EndoVis18	seq2	681
EndoVis18	seq5	335
EndoVis18	seq9	262
EndoVis18	seq15	263
EndoVis17	te1	149
EndoVis17	te2	153
EndoVis17	te3	126
EndoVis17	te4	87
EndoVis17	te5	50
EndoVis17	te6	84
EndoVis17	te7	54
EndoVis17	te8	112
EndoVis17	te9	436
EndoVis17	te10	466

B QA CONTRACT AND METRIC DEFINITIONS

B.1 PROMPT CONSTRUCTION

The six tasks use the input and output schemas below. Model-specific system prompts, chat templates, and multi-turn wrappers form additional parts of the inference input.

- **Depth:** image and specified tip location; output optical-axis depth in millimeters.
- **Inter-instrument distance:** image and two specified instrument locations; output their reconstructed 3D separation.
- **Instrument-to-tissue distance estimation:** image with both specified instrument and selected tissue pixels; output their reference 3D separation.
- **Camera motion:** prescribed image pair or sequence; output translation and rotation in the benchmark convention.
- **Tracking:** observed image sequence and specified instrument referents; output the corresponding structured 3D path.
- **Instrument path completion:** first and last images plus start/end relative 3D coordinates; output seven points including the endpoints.

Tracking supplies seven observed images and requests seven relative 3D points. Completion supplies the first and last images and requests seven points including the endpoints. Frame intervals are specified by each sample. Coordinates and distances are in millimeters unless otherwise stated.

B.2 SCALAR METRICS

For a scored scalar prediction $\hat{y}_i$ and reference y_i, let $e_i = \hat{y}_i - y_i$. On the scored set $\mathcal{V}$ with $n = |\mathcal{V}|$, the standard scalar measures are

$$\mathrm{MAE} = \frac{1}{n}\sum_{i\in\mathcal{V}} |e_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i\in\mathcal{V}} e_i^2}. \tag{1}$$

The scorer uses population standard deviation of absolute error for σ_1 (mm), and population standard deviation of $|e_i|/|y_i|$ multiplied by 100 for σ_2 (percent). Zero-valued references are excluded only from the latter. Invalid or unmatched predictions are excluded and reported in coverage counts. Scalar values are extracted only from accepted answer blocks, as specified below.

B.3 CAMERA METRICS

Let $a_i = \|\hat{\mathbf{t}}_i\|_2 - \|\mathbf{t}_i\|_2$ and $b_i = \|\hat{\mathbf{r}}_{i,\mathrm{Euler}}\|_2 - \|\mathbf{r}_{i,\mathrm{Euler}}\|_2$. Camera metrics are

$$\mathrm{T{-}RMS} = \sqrt{\mathbb{E}[a_i^2]}, \quad \mathrm{E{-}RMS} = \sqrt{\mathbb{E}[b_i^2]}, \quad \mathrm{T{-}MAE} = \mathbb{E}[|a_i|]. \tag{2}$$

These metrics compare translation-vector and Euler-vector magnitudes. The parser requires a complete structured answer containing translation and rotation triples, or a standalone ordered six/seven-number tuple in the requested units. Metrics are computed without global alignment.

B.4 TRACKING AND COMPLETION METRICS

Accept only complete trajectories of exactly seven finite 3D points and exclude the supplied initial point from point-error metrics. MPTE and MeanDev pool Euclidean errors over the remaining points; PCK@2 mm is the pooled percentage strictly below 2 mm. LenDev is the mean absolute difference between predicted and reference path lengths over complete seven-point paths. Incomplete and extra-point paths are rejected; valid-path and point counts accompany the evaluation records.

Final-answer extraction and failures. The same parser is used for every model without access to reference values or model identities. Closed reasoning spans are removed; an unclosed reasoning span is rejected. Explicit answer/final segments take precedence. Otherwise a scalar must occupy

the complete response or terminal answer paragraph; labeled scalar values and single-quantity measurement statements are accepted, with scientific notation and mm/cm/m conversion (omitted units mean the requested mm). Arbitrary reasoning numbers, ranges, placeholders, and tool requests are not scalar answers. Camera and trajectory outputs may occupy a unique standalone structured paragraph or code block. Distinct competing structured candidates are rejected rather than chosen using reference error. Standard labels and markup are normalized. Seven-point trajectories are neither padded nor truncated. Supplied origin values are not corrected by the parser.

Input-unmatched, upstream-error, and parsing-failure records remain in the coverage records. They do not receive zero error and are excluded from conditional error summaries; counts and parse outcomes accompany every result. The deterministic parser applies the same output-format rules across models. Format validation covers finite values, units, camera schemas, and trajectory completeness.

For Group-RMS, first average point errors within each trajectory, group these means by the metadata instrument label, compute RMS within each group, then average the group RMS values with equal group weights. This definition is shared by all seven VLMs and Surgical SAM 2. All tracking methods require seven finite points and score the six noninitial points.

C Source Characterization and Reference Diagnostics

C.1 Label classes and instrument footprint

EndoVis17 lists six evaluated instrument types: Bipolar Forceps, Prograsp Forceps, Large Needle Driver, Monopolar Curved Scissors, Vessel Sealer, and Grasping Retractor, with two to three types per sequence. EndoVis18 lists seven instrument-related labels (shaft, clasper, wrist, clamps, thread, suction instrument, ultrasound probe) and three anatomical labels (kidney parenchyma, covered kidney, intestine), with five to eight classes per sequence. These are different label ontologies. StereoMIS does not provide a comparable anatomy-class inventory in the reported characterization.

Table 10: Instrument-pixel fraction on the reported 1,000 query instances. This is a visibility proxy, not measured occlusion frequency.

Dataset	N	Mean	Median	Std	Min	Max	Below 5%
EndoVis18	300	14.9%	15.0%	5.6%	2.0%	34.8%	2.7%
EndoVis17	300	6.9%	6.1%	3.2%	0.4%	15.1%	33.0%
StereoMIS	400	9.6%	7.3%	7.9%	0.01%	42.6%	18.0%

Table 11: Heuristic uncertainty at the 1,000 depth-query pixels. Relative uncertainty is the stored heuristic sigma divided by raw positive depth (percent); valid denominators are listed.

Dataset	Queries	Below 0.4	Valid σ/Z	Median (%)	95th pct. (%)
EndoVis18	300	114	295	1.051	9.732
EndoVis17	300	105	292	0.557	5.806
StereoMIS	400	310	334	4.191	14.214

Table 12: Sensitivity of the below-threshold fraction to descriptive confidence cutoffs. All columns use the same 1,000 final depth-query pixels and stored maps; values are counts (percent).

Dataset	Queries	$c < 0.2$	$c < 0.3$	$c < 0.4$
EndoVis18	300	101 (33.67%)	109 (36.33%)	114 (38.00%)
EndoVis17	300	98 (32.67%)	100 (33.33%)	105 (35.00%)
StereoMIS	400	306 (76.50%)	307 (76.75%)	310 (77.50%)
All sources	1000	505 (50.50%)	516 (51.60%)	529 (52.90%)

C.2 Interpretation of confidence

We characterize stored confidence at all 1,000 depth-query pixels using cutoffs of 0.2, 0.3, and 0.4 (Table 12). The pooled fractions below these cutoffs are 50.5%, 51.6%, and 52.9%, respectively. At 0.4, the fractions are 114/300 (38.0%) for EndoVis18, 105/300 (35.0%) for EndoVis17, and 310/400 (77.5%) for StereoMIS. Within the corresponding subsets with confidence ≥ 0.4 (186, 195, and 90 queries), heuristic relative-uncertainty medians are 0.536%, 0.370%, and 0.368%, with 95th percentiles of 2.734%, 2.182%, and 2.778%. Relative uncertainty is computed from the stored uncertainty and positive depth at the same raw-depth stage. These statistics characterize the confidence distribution and its relationship to heuristic uncertainty.

C.3 Annotation representation consistency

Serialized answers agree with their corresponding metadata within the stated decimal-rounding tolerance for all 5,998 QA instances. Table 13 gives task-level counts. All 2,000 trajectory annotations contain seven finite 3D points, seven frame identifiers, and a zero relative origin. Projection checks on 3,996 annotated query points agree with the stored pixel coordinates up to floating-point precision when intrinsics are scaled to image resolution. These checks assess numerical representation and internal geometric consistency.

Table 13: Answer–metadata agreement within decimal-rounding tolerance.

Task	QA instances	Within tolerance
Depth	1000	1000
Inter-instrument distance	998	998
Tissue distance	1000	1000
Camera motion	1000	1000
Tracking	1000	1000
Completion	1000	1000

D PER-SOURCE EXPERIMENTAL RESULTS

These source-specific scores and the main aggregate tables are generated from the same input-matched per-example records. N is the number of successfully parsed QA records. Unmatched inputs and failed parses are excluded, not scored as zero. Point-weighted trajectory means and path-weighted length errors use their respective valid denominators.

Table 14: Valid scored QA counts. Nominal task sizes are 1,000, except inter-instrument distance (998). These are conditional scores with model-dependent coverage.

Model	Depth	Dist.	Tissue	Camera	Track	Compl.
GPT-5.5	1000	998	265	999	927	679
Qwen3.7-Plus	1000	998	265	1000	983	666
Kimi-K2.6	949	964	262	954	597	668
Claude-Opus-4.8	1000	995	261	1000	992	678
Gemini-3.5-Flash	1000	969	265	1000	856	671
ViLaSR	328	651	237	20	84	269
VST-7B-RL	1000	998	620	481	671	362

D.1 INSTRUMENT ABSOLUTE DEPTH

Table 15: Instrument absolute depth on EndoVis18.

Model	N	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
GPT-5.5	300	21.71	60.27	56.22	19.93
Qwen3.7-Plus	300	19.67	60.11	56.80	19.12
Kimi-K2.6	260	53.78	132.02	120.58	202.37
Claude-Opus-4.8	300	30.61	69.30	62.17	57.11
Gemini-3.5-Flash	300	**19.54**	**56.65**	**53.18**	20.15
ViLaSR	79	60.25	122.84	107.05	22.94
VST-7B-RL	300	70.01	91.05	58.21	**0.35**

Table 16: Instrument absolute depth on EndoVis17.

Model	N	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
GPT-5.5	300	41.47	61.29	45.13	45.41
Qwen3.7-Plus	300	26.96	52.70	45.28	55.65
Kimi-K2.6	299	57.83	133.06	119.84	178.11
Claude-Opus-4.8	300	45.78	101.91	91.04	129.19
Gemini-3.5-Flash	300	**22.69**	**44.22**	**37.95**	23.69
ViLaSR	126	42.70	74.48	61.03	26.34
VST-7B-RL	300	70.25	80.36	39.02	**2.95**

Table 17: Instrument absolute depth on StereoMIS.

Model	N	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
GPT-5.5	400	44.85	68.25	51.45	186.19
Qwen3.7-Plus	400	36.91	59.21	46.29	127.06
Kimi-K2.6	390	47.97	81.76	66.20	174.03
Claude-Opus-4.8	400	81.70	118.17	85.38	234.89
Gemini-3.5-Flash	400	39.01	58.21	43.20	124.29
ViLaSR	123	**21.94**	**29.18**	**19.24**	76.26
VST-7B-RL	400	48.53	71.22	52.13	**1.63**

D.2 Inter-instrument distance

Table 18: Inter-instrument distance on EndoVis18.

Model	N	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
GPT-5.5	41	27.39	45.23	36.00	34.65
Qwen3.7-Plus	41	39.38	48.38	28.10	56.90
Kimi-K2.6	41	35.86	55.30	42.10	26.32
Claude-Opus-4.8	41	28.04	45.49	35.82	41.34
Gemini-3.5-Flash	41	30.26	43.02	30.57	46.92
ViLaSR	11	**21.61**	**23.09**	**8.14**	17.85
VST-7B-RL	41	57.68	74.03	46.41	**0.54**

Table 19: Inter-instrument distance on EndoVis17.

Model	N	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
GPT-5.5	217	51.56	58.72	28.10	85.33
Qwen3.7-Plus	217	60.47	68.21	31.56	102.06
Kimi-K2.6	213	28.05	34.66	20.36	74.85
Claude-Opus-4.8	217	34.04	37.34	15.33	55.10
Gemini-3.5-Flash	210	56.98	63.69	28.45	106.40
ViLaSR	209	**27.89**	37.47	25.02	87.54
VST-7B-RL	217	28.61	**29.74**	**8.14**	**1.87**

Table 20: Inter-instrument distance on StereoMIS.

Model	N	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
GPT-5.5	740	50.43	103.52	90.41	179.12
Qwen3.7-Plus	740	53.10	102.16	87.27	138.21
Kimi-K2.6	710	40.98	102.23	93.65	159.79
Claude-Opus-4.8	737	40.78	97.97	89.07	149.17
Gemini-3.5-Flash	718	55.68	103.53	87.29	246.41
ViLaSR	431	**31.78**	**79.06**	**72.39**	219.07
VST-7B-RL	740	41.56	105.79	97.29	**3.33**

D.3 INSTRUMENT-TO-TISSUE DISTANCE ESTIMATION

Table 21: Instrument-to-tissue distance estimation on EndoVis18.

Model	N	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
GPT-5.5	141	**0.37**	**0.65**	**0.54**	44.55
Qwen3.7-Plus	141	1.01	1.20	0.65	318.53
Kimi-K2.6	139	1.20	1.59	1.04	435.27
Claude-Opus-4.8	138	3.88	4.77	2.76	2320.88
Gemini-3.5-Flash	141	1.11	1.59	1.14	960.06
ViLaSR	118	1.62	2.39	1.76	972.32
VST-7B-RL	300	0.83	1.12	0.74	**17.99**

Table 22: Instrument-to-tissue distance estimation on EndoVis17.

Model	N	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
GPT-5.5	78	0.97	1.91	1.65	67.17
Qwen3.7-Plus	78	1.33	1.71	1.07	258.64
Kimi-K2.6	77	1.17	**1.46**	0.87	337.92
Claude-Opus-4.8	77	5.64	6.87	3.94	6437.31
Gemini-3.5-Flash	78	2.43	4.77	4.10	838.36
ViLaSR	77	1.80	1.95	**0.74**	798.44
VST-7B-RL	300	**0.81**	1.81	1.61	**16.55**

Table 23: Instrument-to-tissue distance estimation on StereoMIS.

Model	N	MAE ↓	RMSE ↓	σ_1 ↓	σ_2 ↓
GPT-5.5	46	18.60	68.55	65.98	77.16
Qwen3.7-Plus	46	18.51	68.43	65.88	603.74
Kimi-K2.6	46	18.57	68.20	65.63	1059.54
Claude-Opus-4.8	46	20.03	68.13	65.12	8171.54
Gemini-3.5-Flash	46	18.93	67.95	65.26	713.97
ViLaSR	42	20.22	71.79	68.89	1196.91
VST-7B-RL	20	**2.51**	**3.86**	**2.93**	**15.26**

D.4 CAMERA EGO-MOTION

Table 24: Camera ego-motion on EndoVis18.

Model	N	T-RMS ↓	E-RMS ↓	T-MAE ↓
GPT-5.5	300	9.55	7.92	3.56
Qwen3.7-Plus	300	10.24	7.53	5.26
Kimi-K2.6	298	9.85	8.04	4.05
Claude-Opus-4.8	300	11.38	7.61	8.53
Gemini-3.5-Flash	300	9.91	8.32	4.01
ViLaSR	5	22.88	**3.18**	15.20
VST-7B-RL	174	**3.92**	4.11	**2.29**

Table 25: Camera ego-motion on EndoVis17.

Model	N	T-RMS ↓	E-RMS ↓	T-MAE ↓
GPT-5.5	299	6.37	6.34	2.29
Qwen3.7-Plus	300	11.60	7.04	5.31
Kimi-K2.6	269	8.28	6.86	3.79
Claude-Opus-4.8	300	11.52	6.34	8.58
Gemini-3.5-Flash	300	6.74	6.46	2.66
ViLaSR	8	41.64	4.89	20.11
VST-7B-RL	161	**2.73**	**2.62**	**1.76**

Table 26: Camera ego-motion on StereoMIS.

Model	N	T-RMS ↓	E-RMS ↓	T-MAE ↓
GPT-5.5	400	0.79	**0.90**	0.33
Qwen3.7-Plus	400	1.98	1.14	0.78
Kimi-K2.6	387	1.22	1.07	0.57
Claude-Opus-4.8	400	11.20	2.30	10.43
Gemini-3.5-Flash	400	0.47	1.05	**0.23**
ViLaSR	7	131.94	10.69	80.33
VST-7B-RL	146	**0.35**	1.27	0.26

D.5 INSTRUMENT PATH TRACKING

Table 27: Instrument path tracking on EndoVis18.

Model	N	MPTE ↓	PCK (%) ↑	Group-RMS ↓
GPT-5.5	296	**19.06**	2.03	**37.75**
Qwen3.7-Plus	296	165.47	0.90	336.18
Kimi-K2.6	178	240.40	0.84	423.17
Claude-Opus-4.8	300	24.31	1.11	40.84
Gemini-3.5-Flash	264	22.20	**2.34**	40.93
ViLaSR	20	521.02	0.00	558.93
VST-7B-RL	184	48.62	1.54	154.80

Table 28: Instrument path tracking on EndoVis17.

Model	N	MPTE ↓	PCK (%) ↑	Group-RMS ↓
GPT-5.5	271	**24.80**	7.13	**34.83**
Qwen3.7-Plus	295	127.49	4.86	269.79
Kimi-K2.6	223	552.19	1.94	794.21
Claude-Opus-4.8	293	24.86	2.67	34.93
Gemini-3.5-Flash	249	28.05	8.30	45.54
ViLaSR	28	555.48	0.60	626.43
VST-7B-RL	204	40.29	**10.78**	78.83

Table 29: Instrument path tracking on StereoMIS.

Model	N	MPTE ↓	PCK (%) ↑	Group-RMS ↓
GPT-5.5	360	46.12	27.96	115.74
Qwen3.7-Plus	392	266.54	23.04	560.00
Kimi-K2.6	196	408.06	13.10	525.31
Claude-Opus-4.8	399	49.24	3.88	112.99
Gemini-3.5-Flash	343	**44.87**	**28.67**	**107.23**
ViLaSR	36	492.13	1.39	540.09
VST-7B-RL	283	387.34	7.89	362.83

D.6 INSTRUMENT PATH COMPLETION

Table 30: Instrument path completion on EndoVis18.

Model	N	MeanDev ↓	LenDev ↓
GPT-5.5	300	**13.15**	69.56
Qwen3.7-Plus	298	13.22	69.84
Kimi-K2.6	298	13.25	69.85
Claude-Opus-4.8	299	13.25	69.60
Gemini-3.5-Flash	296	13.25	**69.26**
ViLaSR	138	28.89	82.37
VST-7B-RL	175	21.54	93.88

Table 31: Instrument path completion on EndoVis17.

Model	N	MeanDev ↓	LenDev ↓
GPT-5.5	300	14.58	**85.70**
Qwen3.7-Plus	290	14.63	85.85
Kimi-K2.6	293	**14.53**	86.86
Claude-Opus-4.8	300	14.54	85.74
Gemini-3.5-Flash	297	14.64	86.18
ViLaSR	95	31.78	98.57
VST-7B-RL	142	24.18	115.84

Table 32: Instrument path completion on StereoMIS.

Model	N	MeanDev ↓	LenDev ↓
GPT-5.5	79	61.41	292.80
Qwen3.7-Plus	78	61.24	291.58
Kimi-K2.6	77	62.57	**289.08**
Claude-Opus-4.8	79	**61.23**	292.90
Gemini-3.5-Flash	78	61.94	295.84
ViLaSR	36	159.26	324.77
VST-7B-RL	45	86.30	312.54

E GEOMETRY-ASSISTED BASELINE SETTINGS

The Endo3R and Surgical SAM 2 implementations specify the settings below. The methods receive auxiliary geometric information and therefore have different input conditions from the VLM baselines.

Endo3R. The inference script specifies `endo3r.pth` and the adapter loads a `DUSt3R_ViTLarge_BaseDecoder_512_dpt.pth` base checkpoint. The default input resolution is 224, with 320 and 512 also supported. The forward call enables evaluation mode and uncertainty checking. Image coordinates are transformed through the adapter's crop-and-resize operations before point extraction.

To convert the reconstruction to metric units, the adapter resizes the reference depth map to the predicted depth resolution using nearest-neighbor interpolation. Over pixels with finite positive depth in both maps, it computes the ratio of reference and predicted depth medians, requiring at least 100 valid pixels. This factor scales depth, distance, and translation predictions. Camera translation uses the first frame's scale factor; rotation is not rescaled. Reference depth supplies the metric scale for this baseline.

Surgical SAM 2. The script specifies the SAM 2.1 Hiera-S configuration `sam2.1_hiera_s.yaml` and an EndoVis18-adapted checkpoint named `sam2.1_hiera_s_endo18.pth`. The first-frame instrument-tip coordinate supplies a positive point prompt. Video mask propagation thresholds mask logits at zero. The propagated masks are combined with stored stereo depth and per-sequence camera calibration to extract three-dimensional tips; candidate tips are associated by proximity to the preceding two-dimensional tip.

Depth refinement uses the median of finite positive depths in a square neighborhood of radius five pixels (up to 11×11), retaining pixels with stored confidence at least 0.5 when a confidence map is available. Without a confidence map, confidence is set to one for this computation. If no valid neighborhood depth remains, the adapter falls back to the geometric tip extractor's depth. When a subsequent frame has no detected tip, the preceding estimate is retained; failure to obtain an initial three-dimensional tip rejects the sample. Trajectories are expressed relative to the first estimated point. The 0.5 threshold applies to this baseline's depth-refinement operation.

E.1 SHARED TRACKING EVALUATION

The 1,000 SAM 2 output records match the evaluation manifest by sample ID, source, sequence, ordered frame IDs, and instrument label; their embedded reference trajectories agree with the reference metadata. Available initial prompt coordinates also agree. Of these records, 23 have pipeline failures and 45 lack a complete finite seven-point trajectory, leaving 932 scored trajectories. MPTE and PCK@2 mm use the 5,592 noninitial points. Group-RMS first averages the six point errors within each trajectory, computes the RMS of these means within each instrument-label group, and then averages those group RMS values equally. The aggregate includes 818 `instrument_0`, 55 `right_instrument`, and 59 `left_instrument` trajectories. Source-specific results use the same definition within each source.

Table 33: Surgical SAM 2 tracking results under the shared metric definitions. Scores are conditional on complete valid trajectories. Absolute errors are mm and PCK is percent.

Source	Valid QA	Points	MPTE ↓	PCK@2mm ↑	Group-RMS ↓
EndoVis18	290	1,740	19.89	14.14	42.57
EndoVis17	261	1,566	30.94	16.48	48.92
StereoMIS	381	2,286	57.14	28.57	141.06
All sources	932	5,592	38.22	20.69	108.22