Decoupling Vision and Reasoning: A Data-Efficient Pipeline for Surgical VQA

Mohamed Hamdy[1] MM1905748@QU.EDU.QA
Fatmaelzahraa Ali Ahmed[2] FATMAAHMED.HMC@GMAIL.COM
Muraam Abdel-Ghani[2] MURAAM.ABDELGHANI@OUTLOOK.COM
Muhammad Arsalan[1] MUHAMMAD.ARSALAN@QU.EDU.QA
Ponnuthurai Nagaratnam Suganthan[1] P.N.SUGANTHAN@QU.EDU.QA
Khalid Al-Jalham[2] KALJALHAM@HAMAD.QA
Abdulaziz Al-Ali[1] A.ALALI@QU.EDU.QA
Shidin Balakrishnan[2] SBALAKRISHNAN1@HAMAD.QA
[1] *Computer Science and Engineering Department, College of Engineering, Qatar University, Doha, Qatar*
[2] *Department of Surgery, Hamad Medical Corporation, Doha, Qatar*

Editors: Under Review for MIDL 2026

Abstract

Vision-language models (VLMs) are becoming increasingly important for surgical intelligence, where reliable scene understanding requires combining visual perception with language-based reasoning. However, progress is constrained by the scarcity of high-quality multimodal datasets, making end-to-end training more prone to overfitting. Existing approaches often address this limitation by converting task-specific datasets (e.g., segmentation, phase recognition, tool-tissue interaction) into synthetic vision-question answering (VQA) form, but such conversions provide only sparse supervision and limit generalization. To overcome these challenges, we propose a modular pipeline that decouples vision information extraction from reasoning. Specialist surgical models–proven effective for their corresponding vision tasks–are first used to extract task-relevant signals, which are then transformed via heuristics into structured textual descriptions. These descriptions, together with the clinical question, are passed to a large language model (LLM) that performs the reasoning step and provides the answer. We evaluate this pipeline on the EndoVis-18-VQA benchmark under different configurations of specialist models and LLMs, showing that combining complementary experts yields stronger performance than relying on any single model. Our approach achieves higher accuracy, recall and F1 than existing surgical VQA baselines, with improvements of up to 2.3% in accuracy without requiring multimodal training, establishing abstraction-driven modularity as a data-efficient and generalizable paradigm for surgical vision-language understanding.

Keywords: Surgical VQA, Modular Vision-Language Models, Vision Language Models, Multi-modal Reasoning

1. Introduction

Minimally invasive surgery now underpin contemporary surgical practice due to their reduced incision size, postoperative pain, and recovery time (Sijberden et al., 2025). These procedures generate continuous, high-resolution endoscopic video that captures anatomy, instruments, and workflow, providing a rich substrate for computer vision and vision-language

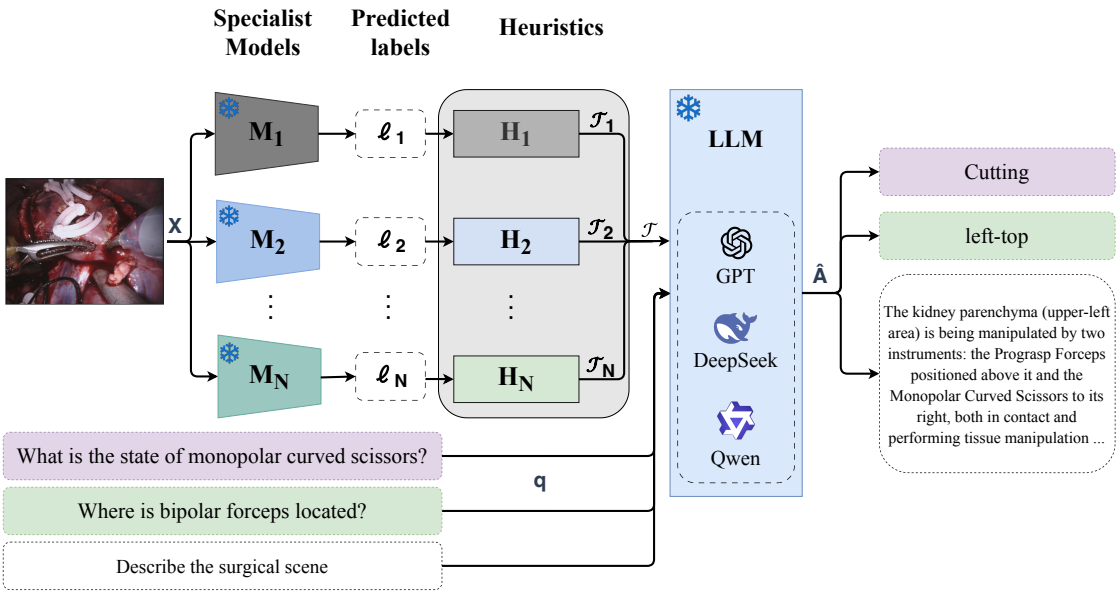

Figure 1: Overview of the proposed modular pipeline. The snowflake denotes frozen models.

methods (Buia et al., 2015; Mascagni et al., 2022). Accordingly, much of the progress in surgical AI has been driven by task-specific vision models for tool and tissue segmentation (Allan et al., 2020; Hong et al., 2020), workflow phase recognition (Twinanda et al., 2016), tool-tissue interaction analysis (Islam et al., 2020; Seenivasan et al., 2022b), and higher-level action triplet recognition (Nwoye et al., 2022).

Despite their success, these models remain constrained by narrowly defined tasks such as segmentation, phase recognition, or interaction classification. To achieve broader reasoning capabilities, recent work has turned toward vision-language models (VLMs), which couple surgical perception with natural-language supervision. By aligning video with text, VLMs offer richer semantics, interpretable outputs, and more flexible clinical interaction. Large-scale efforts such as SurgVLM (Zeng et al., 2025) and SurgVLP (Yuan et al., 2025) illustrate the feasibility of training generalist surgical models, while systems like Surgical-GPT (Seenivasan et al., 2023) and Surgical-VQA (Seenivasan et al., 2022a; Yuan et al., 2024a) highlight the potential of answering clinically relevant queries directly from surgical scenes.

A central challenge for surgical vision-language modeling is the dependence on large multimodal datasets, which remain scarce and costly to curate (Yuan et al., 2024b). Unlike natural-image domains with abundant paired corpora, surgical data requires expert annotation and is often limited to specific procedures. To expand supervision, prior works repurpose existing task-specific datasets—such as segmentation, workflow recognition, or tool-tissue interaction—into synthetic VQA form (Seenivasan et al., 2022a; Yuan et al., 2024a; Zeng et al., 2025). While effective, these conversions provide only narrow language supervision tied to single queries (Wang et al., 2024b; Menon and Vondrick, 2022) and can

introduce noise when relying on automatically transcribed audio (Yuan et al., 2024c,b). These limitations motivate frameworks that reduce reliance on synthetic multimodal alignment and instead directly leverage existing expert-trained models.

In this work, we propose a modular pipeline for surgical vision-language understanding that overcomes these limitations. Instead of training end-to-end multimodal models, we decouple perception from reasoning by relying on frozen, expert-trained vision models to extract task-specific information, which is then translated into structured textual descriptions for language-based inference. This separation offers two main advantages. First, it eliminates the need for large-scale multimodal supervision, avoiding both the inefficiency of synthetic VQA conversions and the noise of weakly aligned transcripts. By transforming visual predictions into human-readable text, the intermediate representations remain interpretable while reducing the risk of overfitting. Second, by incorporating multiple specialized models–such as segmentation and tool-tissue interaction networks–we aggregate complementary signals that together provide a richer and more clinically relevant description of the scene.

Our contributions are three-fold:

- We introduce a general framework that abstracts surgical perception into textual descriptions, enabling downstream reasoning without multimodal training.
- We define heuristics to convert outputs from specialist models into interpretable inputs for large language models, making the reasoning process transparent.
- We demonstrate through extensive experiments on EndoVis-18-VQA that combining complementary vision experts with LLM reasoning surpasses prior baselines, establishing abstraction-driven modularity as a data-efficient and generalizable paradigm for surgical VQA.

2. Problem Setup and Notation

We study the problem of surgical visual question answering (VQA), where the input is an endoscopic frame and a natural language question, and the goal is to return a clinically relevant answer. Our design departs from traditional end-to-end vision–language models by decoupling perception from reasoning. Specifically, we factor the pipeline into three stages:(i) task-specific *specialist models* that extract visual information, (ii) *heuristics* that map these predictions into structured textual representations, and (iii) a frozen large language model (LLM) that performs reasoning. This modular decomposition reduces the dependence on large-scale multimodal datasets, while maintaining interpretability and flexibility.

Inputs. Let $X \in \mathbb{R}^{H\times W\times 3}$ denote a surgical image, and let q denote a natural language question. The objective is to predict an answer $\hat{A}$ expressed in natural language.

Specialist models. Previous research has shown that a single vision encoder typically captures only part of the information contained in visual scenes. For example, language-supervised encoders such as CLIP (Radford et al., 2021) align well with semantics but often miss fine-grained spatial cues, whereas self-supervised encoders like DINOv2 (Oquab et al., 2023) or segmentation experts like SAM (Kirillov et al., 2023) capture complementary

information (Fan et al., 2024; Tong et al., 2024). Multi-expert systems that combine such models consistently improve representation quality and downstream performance (Shi et al., 2024). Motivated by these insights, we assume access to a collection of N task-specific vision experts $\{M_i\}_{i=1}^{N}$, each trained to solve a distinct perception problem. Given an input image X, each expert produces a task-level prediction

$$\ell_i = M_i(X),$$

where the form of ℓ_i depends on the task—for instance, a categorical label, a dense segmentation mask, or a set of interaction triplets.

Heuristics. The outputs ℓ_i produced by the specialist models are heterogeneous, ranging from discrete labels to dense masks or structured tuples. Multimodal learning often fuses such outputs via trainable projection layers, requiring large-scale multimodal datasets, which are scarce in surgical domains. To avoid these limitations, we introduce a family of deterministic heuristics that abstract model predictions into symbolic textual descriptions, thus enabling reasoning entirely in the language domain while keeping the LLM frozen. Formally, for each task i we define

$$H_i : \ell_i \mapsto \mathcal{T}_i,$$

where $\mathcal{T}_i$ is a set of textual statements describing ℓ_i. Aggregating across all models yields

$$\mathcal{T} = \bigcup_{i=1}^{N} \mathcal{T}_i,$$

an intermediate representation for reasoning. These heuristics range from simple templates (e.g., interaction labels) to more complex rule-based procedures (e.g., spatial relations from segmentation masks).

Language reasoning. Once visual predictions are abstracted into text, the reasoning step is handled by an LLM. Unlike traditional multimodal systems requiring joint training to align image and text embeddings, we use a frozen LLM purely as an inference engine. This avoids additional multimodal supervision and leverages the broad reasoning capabilities of pretrained language models. Formally, given a frozen LLM, q and $\mathcal{T}$ derived from heuristics, the model predicts an answer

$$\hat{A} = \mathrm{LLM}(q, \mathcal{T}).$$

Operating exclusively on interpretable textual descriptions provides a flexible means of combining heterogeneous visual cues without requiring any gradient-based adaptation to the surgical domain as the perception burden is already addressed by the specialist models.

Overall pipeline. The complete formulation can thus be expressed as

$$X \xrightarrow{M_i} \ell_i \xrightarrow{H_i} \mathcal{T}_i \xrightarrow{\cup} \mathcal{T} \xrightarrow{\mathrm{LLM}(q,\cdot)} \hat{A}.$$

An illustration of this general pipeline is provided in Figure 1.

3. Methodology

Our methodology instantiates the general pipeline described in Section 2 for the task of surgical VQA. We detail the concrete choices of specialist models, the design of heuristics for textual abstraction, and the language reasoning stage.

3.1. Specialist Model Choices

To cover essential aspects of surgical scene understanding, we rely on two classes of vision experts: *segmentation models* and *interaction models*, capturing complementary low-level spatial structure and high-level functional cues.

3.1.1. Segmentation

Segmentation provides dense spatial evidence–what structures are present, where they are located, and how elements are arranged. Without such structure, an LLM lacks the spatial grounding needed for clinically precise answers. Prior work suggests that segmentation masks alone can support the generation of clinically relevant descriptions of surgical scenes (Hamdy et al., 2025). Segmentation therefore supplies a dense, procedure-agnostic representation $S \in \{0, \ldots, C\}^{H \times W}$ that our heuristics convert into textual facts. We experiment with two models: FASL and a hybrid SAM+FASL variant.

FASL. The Feature-Adaptive Spatial Localization model (FASL) (Abdel-Ghani et al., 2025) improves holistic surgical scene segmentation by combining low-level and high-level features. It achieves strong performance across instrument and anatomy segmentation benchmarks, surpassing prior baselines. These capabilities make FASL a suitable choice for providing the semantic scene information required by our pipeline.

SAM+FASL. While specialist pixel-wise classifiers perform strongly on curated datasets, they may struggle under distribution shift or produce fragmented labels. Prior work suggests decoupling *mask generation* from *classification* via foundation segmenters followed by domain-specific classifiers (Hu et al., 2023; Wang et al., 2024c). Following this strategy, we construct a hybrid SAM+FASL model: SAM (Kirillov et al., 2023) produces region masks $\mathcal{R} = \{R_k\}$, while FASL generates a label map S. Each region is assigned the class

$$\hat{y}(R_k) \;=\; \arg\max_{c \in \{1,\ldots,C\}} \big|\{\, p \in R_k \;\mid\; S(p) = c\}\big|,$$

that is, the class most frequently predicted by FASL within the region. This majority-vote labeling leverages SAM's generalization for region proposals while retaining FASL's domain-specific semantics for class assignment. A qualitative example of this process is shown in Appendix Figure 3.

3.1.2. Interaction

Understanding tool-tissue interactions is a key step in surgical scene understanding, since it provides not only knowledge of *what* instruments and anatomical structures are present, but also *how* they functionally relate during a procedure, which is important for skill assessment, feedback, and decision support (Islam et al., 2020).

Tool-tissue interaction. We adopt the graph-based interaction model of (Seenivasan et al., 2022b), where instruments and tissues are nodes and functional relationships (e.g., grasping, cutting, retracting, suction, idle) are edges. Using both visual-semantic and relational reasoning, the model predicts the interaction state of each instrument-tissue pair.

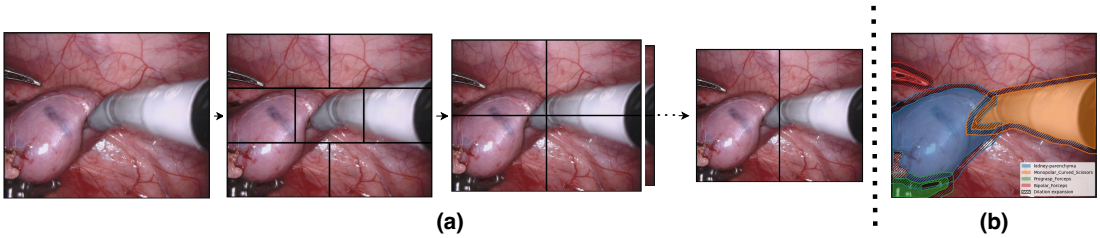

Figure 2: Segmentation heuristics: (a) absolute location via prioritized spatial regions; (b) dilation-based adjacency.

3.2. Heuristics Design

To instantiate the abstraction step, we apply deterministic heuristics that convert the heterogeneous predictions of the specialist models into structured textual descriptions. These rules bridge perception and reasoning by translating segmentation masks and interaction labels into interpretable, token-efficient statements that the LLM can consume.

3.2.1. Segmentation

Given a semantic segmentation mask S, the module H_{seg} produces a fact set summarizing object presence, coarse layout, and spatial relationships. A high-level illustration of this process is shown in Figure 2. The module generates three families of statements:

Absolute location. Each object instance is assigned to a symbolic region (e.g., "top-left", "center") based on a small hierarchy of spatial partitions, enabling a coarse but stable description of where structures appear.

Pairwise spatial relations. When multiple objects occupy nearby regions, we infer simple directional relations (e.g., "the forceps is to the left of the needle") using adjacency cues and centroid orientation.

Graded proximity. For instrument-tissue pairs, we estimate discrete proximity levels (e.g., "touching", "near", "far") through successive neighborhood expansions around the instrument mask, providing soft cues about closeness.

Together, these heuristics yield $\mathcal{T}_{\text{seg}}$, a compact symbolic summary of the scene.

3.2.2. Interaction

The interaction model outputs categorical labels describing functional relations such as coagulation, cutting, or idle. The heuristics module H_{int} converts these predictions into concise declarative facts using simple templates, such as "monopolar curved scissors: cutting," forming the fact set $\mathcal{T}_{\text{int}}$. These statements complement the spatial information extracted from segmentation, enabling the LLM to reason about both *where* objects are and *how* they are being used.

Full algorithmic details for all heuristics are provided in Appendix A.

3.3. Language Reasoning

The fact sets $\mathcal{T}_{\text{seg}}$ and $\mathcal{T}_{\text{int}}$ are concatenated into a unified description $\mathcal{T}$. Given a question q, the frozen LLM predicts the answer $\hat{A}$ by reasoning over this structured textual summary.

4. Experimental Setup

Datasets. We evaluate our pipeline on the test split of the **EndoVis-18-VQA** benchmark (Seenivasan et al., 2022a), which contains roughly 2.7k question–answer pairs. Specialist models are trained separately: the segmentation expert (FASL) uses the EndoVis-18 annotations (Allan et al., 2020), and the tool–tissue interaction expert is trained on the dataset of Islam et al. (2020). To avoid data leakage, videos 1, 5, and 16 are held out for testing across all specialists.

Competing models. We benchmark against representative surgical VQA baselines, including VisualBERT (Li et al., 2019), VisualBERT RM (Seenivasan et al., 2022a), and LV-GPT with RN18 and Swin backbones (Seenivasan et al., 2023). We also compare with Surgical-LVLM (Wang et al., 2024a) and recent surgical vision-language models such as SurgVLM (Zeng et al., 2025).

Language models. For the reasoning stage, we evaluate multiple frozen LLMs accessed via their official APIs. From OpenAI, we include GPT-4o, GPT-4o-mini, GPT-5, GPT-5-mini, and GPT-5-nano; from DeepSeek, we include DeepSeek-chat and DeepSeek-reasoner. We additionally test open-source vision-language models from the Qwen-VL family (2.5 and 3.0), including the 32B, 72B, and 8B instruct variants. All models operate on identical serialized textual fact sets to ensure fair comparison.

Implementation details. Specialist models are trained and evaluated on a workstation with an RTX 4090 GPU and Intel i9 CPU. FASL follows the training protocol of Abdel-Ghani et al. (2025), while the interaction expert is loaded from a publicly available checkpoint.[1] LLMs are queried without further fine-tuning. Evaluation metrics include overall accuracy, recall, and F1 score.

5. Results and Discussion

Specialist models. Table 1 summarizes performance across segmentation-only, interaction-only, and combined specialist configurations. The oracle rows (GT) represent upper bounds obtained by employing ground-truth masks or interaction labels directly. Among single experts, segmentation models clearly dominate: FASL reaches 62% accuracy, compared to 45% for interaction-only predictions, indicating that LLMs can infer a broader range of clinically relevant answers from spatial cues provided by segmentation masks than from interaction labels alone. Combining segmentation and interaction specialists consistently improves performance, with FASL+INT achieving 73% accuracy and 36% F1. Using ground-truth for both tasks (GT+GT) yields 90% accuracy and 83% F1, quantifying the potential performance ceiling if perception errors were eliminated.

The per-question breakdown provides further insight. The GT+GT oracle achieves perfect accuracy on organ and state questions but still incurs occasional errors on spatial

1. https://github.com/lalithjets/Global-reasoned-multi-task-model

Table 1: Performance of segmentation and interaction specialists on EndoVis-18-VQA. Rows are grouped into segmentation-only, interaction-only, and combined configurations. Columns under *Overall Performance* report Acc, Recall, and F1, while those under *Question Type (Acc)* report accuracy per question category (Organ, State, Location). 'GT' denotes oracle annotations (i.e., ground truth provided directly), and 'INT' refers to the tool-tissue interaction specialist model. Bold values indicate the best performance among non-oracle models. All results in the table are reported using GPT-4o-mini as the reasoning LLM.

Seg. Model	Interaction	Overall Performance			Question Type (Acc)		
		Acc	Recall	F1	Organ (16%)	State (42%)	Location (42%)
FASL	—	0.6208	**0.3674**	0.3397	**1.0000**	0.4091	0.6865
SAM+FASL	—	0.6194	0.3354	0.3032	**1.0000**	0.4539	0.6382
GT	—	0.6677	0.5337	0.4570	1.0000	0.4109	0.7967
—	INT	0.4536	0.2039	0.1558	**1.0000**	**0.6916**	0.0052
—	GT	0.5778	0.5770	0.4274	1.0000	0.9905	0.0026
GT	GT	0.8996	0.8242	0.8304	1.0000	0.9543	0.8062
GT	INT	0.7862	0.4243	0.4336	1.0000	0.6693	0.8208
FASL	GT	0.8472	0.7497	0.7523	1.0000	0.9414	0.6942
FASL	INT	**0.7299**	0.3516	**0.3605**	**1.0000**	0.6615	**0.6942**

queries. These mistakes typically stem not from missing visual information but from ambiguous boundaries and the discretized answer space: even when the LLM has enough facts to infer that a tool is located at the center-right, it must choose among predefined coarse regions (e.g., "top-right" or "bottom-right"), sometimes producing an answer that disagrees with the ground-truth tag. Beyond this, the per-question evaluation shows the complementary strengths of the two modalities. Segmentation variants excel at location-oriented questions (FASL: 69% accuracy) but struggle on state questions, whereas interaction models are strongest on state queries (INT: 69%) yet almost fail to capture location (0.5%). Combining both sources recovers the best of each, yielding more balanced performance across categories and confirming that segmentation and interaction specialists provide complementary signals for the necessary reasoning.

Language models. Table 2 compares frozen LLMs across three input settings: specialist predictions (FASL+INT), oracle annotations, and a image-only condition. With specialist inputs, GPT-5 achieves the strongest performance (77.4% accuracy, 47.6% F1), with GPT-5-mini and GPT-5-nano close behind. DeepSeek-chat and DeepSeek-reasoner perform reasonably well but remain several points lower in F1, while GPT-4o/4o-mini and the Qwen-VL models lag further. Under oracle inputs, all models improve substantially: GPT-5-mini reaches the highest scores (93.0% accuracy, 95.3% F1), with GPT-5 and the stronger Qwen-VL variants close behind. The narrow spread between LLMs in the oracle setting highlights that, once provided with perfect scene descriptions, even compact language models can be effectively employed within our modular pipeline.

Table 2: Comparison of LLMs on EndoVis-18-VQA under three input settings: (i) **FASL, INT** uses predictions from both specialists, (ii) **Oracle** uses ground-truth masks and interaction labels, and (iii) **Only Images** provides the raw frame with no specialist input. Bold values indicate the best performance within each setting.

	FASL, INT			Oracle			Only Images		
	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1
deepseek-chat	0.7429	0.4504	0.4551	0.9256	0.9474	0.9367	–	–	–
deepseek-reasoner	0.7591	0.4072	0.4127	0.9296	0.8904	0.8889	–	–	–
gpt-4o	0.7443	0.3985	0.4064	0.9155	0.7135	0.7236	–	–	–
gpt-4o-mini	0.7299	0.3516	0.3605	0.8996	0.8242	0.8304	–	–	–
gpt-5	**0.7739**	**0.4706**	**0.4758**	0.9299	0.9500	0.9481	**0.3723**	**0.2979**	**0.2854**
gpt-5-mini	0.7656	0.4656	0.4723	**0.9303**	**0.9502**	**0.9528**	–	–	–
gpt-5-nano	0.7660	0.4373	0.4447	0.9252	0.8890	0.8608	0.3438	0.2497	0.2136
qwen2.5-vl-32b-instruct	0.7198	0.1818	0.1827	0.9047	0.4671	0.4603	0.2723	0.0593	0.0368
qwen2.5-vl-72b-instruct	0.7562	0.3629	0.3684	0.9260	0.8886	0.8906	0.3604	0.1544	0.1252
qwen3-vl-8b-instruct	0.7244	0.2641	0.2706	0.9075	0.6384	0.6410	0.2626	0.0974	0.0765

Image-only reasoning. The image-only condition in Table 2 evaluates each LLM using only the raw endoscopic frame. This isolates the contribution of our modular pipeline: the same LLM is evaluated once with specialist-derived facts and once with no structured perception. Performance drops sharply across all models (e.g., GPT-5: 37.2% accuracy, 28.5% F1), showing that even strong LLMs struggle to infer clinically precise spatial or semantic cues from pixels alone. The consistent gap between image-only and specialist-assisted performance highlights the importance of the modular abstraction stage in enabling frozen LLMs to operate effectively in surgical settings without multimodal fine-tuning. The consistent gap between image-only and specialist-assisted performance isolates the contribution of the modular abstraction stage in adapting frozen LLMs to the surgical domain without fine-tuning, especially given the scarcity of multimodal surgical datasets.

Baselines. Table 3 summarizes performance on EndoVis-18-VQA. Early transformer-based approaches such as VisualBERT and its residual variant (VisualBERT RM) achieved accuracies around 61–62%, with limited recall and F1, reflecting the difficulty of learning surgical semantics from limited VQA supervision. More recent efforts sought to improve reasoning by augmenting vision features. SurgicalGPT (LV-GPT) introduced a GPT-based architecture with vision token embeddings and careful token sequencing, reaching 66–68% accuracy. Surgical-LVLM further adapted a large vision-language model with Visual Perception LoRA modules, attaining close to 70% accuracy. The most recent large-scale effort, SurgVLM, leveraged instruction tuning on multimodal surgical data and achieved 75.0% accuracy (Zeng et al., 2025).

In contrast, our proposed modular pipeline, which avoids multimodal fine-tuning altogether, surpasses these baselines. With GPT-5-nano, it achieves 76.6% accuracy, while GPT-5 attains 77.4% accuracy with the highest F1 score (47.6%). These results show that structured, specialist-driven scene abstraction combined with frozen LLMs can outperform both earlier transformer models and recent foundation-scale VLMs.

Table 3: Performance comparison between prior surgical VQA approaches and our modular pipeline on EndoVis-18-VQA. ‡ denotes variants belonging to our pipeline.

Model	Acc	Recall	F1
VisualBert (Li et al., 2019)	0.6143	0.4282	0.3745
VisualBert RM (Seenivasan et al., 2022a)	0.6190	0.4079	0.3583
LV-GPT (RN18) (Seenivasan et al., 2023)	0.6811	0.4649	0.4649
LV-GPT (Swin) (Seenivasan et al., 2023)	0.6613	0.4460	0.4537
Surgical-LVLM (Wang et al., 2024a)	0.6947	–	0.3325
SurgVLM-72B Lora-tuning (Zeng et al., 2025)	0.7502	–	–
Ours (gpt-5-nano)‡	0.7660	0.4373	0.4447
Ours (gpt-5)‡	**0.7739**	**0.4706**	**0.4758**

Discussion. Unlike prior approaches that depend on multimodal pretraining or direct VQA-style supervision, our framework achieves superior performance by decoupling perception from reasoning. Models such as VisualBERT RM and Surgical-GPT rely on VQA-style training, whereas SurgVLM leverages large-scale multimodal pretraining. In contrast, our modular abstraction pipeline requires only task-specific training of vision specialists, offering a more data-efficient and generalizable paradigm. The textual intermediate representation provides interpretability, and the modular design enables seamless incorporation of additional experts for new skills or procedures, allowing the pipeline to adapt to new datasets through strongly supervised specialist training rather than expensive multimodal data collection or fine-tuning of the language model.

Limitations. The current evaluation focuses on vision-centric queries in EndoVis-18, which does not capture knowledge-intensive or temporally extended reasoning. While the pipeline is expected to generalize with appropriate specialists, broader validation is essential. The modular design, while flexible, introduces additional computational overhead at inference time since multiple experts must be executed before reasoning. Future work, shall study developing lightweight expert-routing mechanisms that activate only the relevant specialists. Extending the pipeline to temporal experts for video understanding, incorporating surgical-domain LLMs for richer clinical reasoning, and evaluating on larger and more diverse datasets represent natural next steps.

6. Conclusion

In conclusion, abstraction-driven modularity emerges as a viable alternative to end-to-end multimodal training for surgical vision-language understanding. By separating perception from reasoning and channeling expert predictions into structured textual facts, the approach achieves state-of-the-art performance on the Endovis-18-VQA surgical benchmark while requiring no multimodal pretraining. These findings highlight both the promise and the trade-offs of modular abstraction, underscoring its potential as a practical foundation for future extensions in surgical intelligence.

Acknowledgments

Research reported in this publication was supported by the Qatar Research Development and Innovation Council (QRDI) grant number ARG01-0522-230266. Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of Qatar Research Development and Innovation Council.

References

Muraam Abdel-Ghani, Mahmoud Ali, Mohamed Ali, Fatmaelzahraa Ahmed, Mohamed Arsalan, Abdulaziz Al-Ali, and Shidin Balakrishnan. Fasl-seg: Anatomy and tool segmentation of surgical scenes. *arXiv preprint arXiv:2509.06159*, 2025.

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. *arXiv preprint arXiv:2001.11190*, 2020.

Alexander Buia, Florian Stockhausen, and Ernst Hanisch. Laparoscopic surgery: a qualified systematic review. *World journal of methodology*, 5(4):238, 2015.

Xiaoran Fan, Tao Ji, Changhao Jiang, Shuo Li, Senjie Jin, Sirui Song, Junke Wang, Boyang Hong, Lu Chen, Guodong Zheng, et al. Mousi: Poly-visual-expert vision-language models. *arXiv preprint arXiv:2401.17221*, 2024.

Mohamed Hamdy, Fatmaelzahraa Ali Ahmed, Mariam Ahmed, Mohannad Natheef AbuHaweeleh, Muraam Abdel-Ghani, Muhammad Arsalan, Abdulaziz Al-Ali, and Shidin Balakrishnan. Segmentation-informed captioning: A multi-stage pipeline for surgical vision–language dataset generation. In *Medical Imaging with Deep Learning - Short Papers*, 2025. URL https://openreview.net/forum?id=3PVEvBmdxc.

W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020.

Yang Hu, Kelly Caylor, and Anna Boser. Segment-then-classify: Few-shot instance segmentation for environmental remote sensing. In *Proceedings of the NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning: Blending New and Existing Knowledge Systems*, 2023.

Mobarakol Islam, Lalithkumar Seenivasan, Lim Chwee Ming, and Hongliang Ren. Learning and reasoning with the graph structure representation in robotic surgery. In *International conference on medical image computing and computer-assisted intervention*, pages 627–636. Springer, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

Liunian Harold Li, Mark Yatskar, D Yin, CJ Hsieh, and KW Chang. Visualbert: A simple and performant baseline for vision and language. arxiv. *arXiv preprint arXiv:1908.03557*, 10, 2019.

Pietro Mascagni, Deepak Alapatt, Luca Sestini, Maria S Altieri, Amin Madani, Yusuke Watanabe, Adnan Alseidi, Jay A Redan, Sergio Alfieri, Guido Costamagna, et al. Computer vision in surgery: from potential to clinical value. *npj Digital Medicine*, 5(1):163, 2022.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022.

Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 33–43. Springer, 2022a.

Lalithkumar Seenivasan, Sai Mitheran, Mobarakol Islam, and Hongliang Ren. Global-reasoned multi-task learning model for surgical scene understanding. *IEEE Robotics and Automation Letters*, 7(2):3858–3865, 2022b.

Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. Surgicalgpt: end-to-end language-vision gpt for visual question answering in surgery. In *International conference on medical image computing and computer-assisted intervention*, pages 281–290. Springer, 2023.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. In *The Thirteenth International Conference on Learning Representations*, 2024.

Jasper P Sijberden, Christoph Kuemmerli, Francesca Ratti, Mathieu D'Hondt, Robert P Sutcliffe, Roberto I Troisi, Mikhail Efanov, Robert S Fichtinger, Rafael Díaz-Nieto,

Giuseppe M Ettorre, et al. Laparoscopic versus open parenchymal preserving liver resections in the posterosuperior segments (orange segments): a multicentre, single-blind, randomised controlled trial. *The Lancet Regional Health–Europe*, 51, 2025.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

Guankun Wang, Long Bai, Wan Jun Nah, Jie Wang, Zhaoxi Zhang, Zhen Chen, Jinlin Wu, Mobarakol Islam, Hongbin Liu, and Hongliang Ren. Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. *arXiv preprint arXiv:2405.10948*, 2024a.

Hao Wang, Fang Liu, Licheng Jiao, Jiahao Wang, Zehua Hao, Shuo Li, Lingling Li, Puhua Chen, and Xu Liu. Vilt-clip: Video and language tuning clip with multimodal prompt learning and scenario-guided optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5390–5400, 2024b.

Jingying Wang, Haoran Tang, Taylor Kantor, Tandis Soltani, Vitaliy Popov, and Xu Wang. Surgment: segmentation-enabled semantic search and creation of visual question and feedback to support video-based surgery learning. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024c.

Kun Yuan, Manasi Kattel, Joel L Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. Advancing surgical vqa with scene graph knowledge. *International journal of computer assisted radiology and surgery*, 19(7):1409–1417, 2024a.

Kun Yuan, Nassir Navab, Nicolas Padoy, et al. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *Advances in Neural Information Processing Systems*, 37:122952–122983, 2024b.

Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 306–316. Springer, 2024c.

Kun Yuan, Vinkle Srivastav, Tong Yu, Joel L Lavanchy, Jacques Marescaux, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. *Medical Image Analysis*, page 103644, 2025.

Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, Erli Zhang, Junde Wu, Jiaan Zhang, Yuxuan Wang, Chang Han Low, Jian Jiang, Zilong Zheng, et al. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *arXiv preprint arXiv:2506.02555*, 2025.

Appendix A. Heuristics for Textual Abstraction

This appendix provides the full technical details of the deterministic heuristics used to convert segmentation masks and interaction predictions into the textual fact sets consumed by the LLM. These procedures correspond directly to the high-level description in Section 3.2.1.

A.1. Segmentation Heuristics

Given a semantic segmentation mask

$$S \in \{0, \dots, C\}^{H \times W},$$

where each pixel encodes one of C semantic classes, the segmentation heuristics module H_{seg} produces a fact set $\mathcal{T}_{\text{seg}}$ summarizing object presence, layout, and spatial relationships. As in the main paper, three families of statements are emitted: *absolute location*, *pairwise spatial relations*, and *graded proximity*.

Absolute location. The goal of this step is to map each object instance into a coarse symbolic location such as "top-left" or "center." Formally, let

$$O_k = \{(i, j) \mid S_{i,j} = k\}$$

denote the set of pixels belonging to class k. We consider a collection of spatial partitions $\mathcal{P}$, each of which defines regions of the image grid. Importantly, these partitions are applied hierarchically where *fine-grained regions* such as quadrants ("top-left," "bottom-right," etc.) are tested first, and only if no assignment is found do we fall back to *coarser partitions* such as halves ("left," "right," "top," "bottom").

For each region $R \in \mathcal{P}$, we measure the overlap with the object pixels:

$$\frac{|O_k \cap R|}{|O_k|}.$$

If at least $\tau = 0.75$ (75%) of the object lies within R, the object is assigned to that region. If no assignment succeeds at the class level, we decompose O_k into connected components $\{O_{k,r}\}_{r=1}^{m}$ and repeat the test per component, yielding a (possibly multi-tag) set of locations for k. This design gives a single stable tag when the majority of the object lies in a single region, yet captures multi-part layouts when the class appears in multiple regions. Figure 2 (a) shows an illustration of this hierarchical procedure.

Pairwise spatial relations. While absolute positions capture coarse symbolic regions, they do not express relational cues when objects occupy the same area and have irregular shapes. To address this, we define pairwise spatial relations based on dilation-based adjacency.

For each object k, we dilate its mask S_k by ρ pixels,

$$S_k^{(\rho)} = \text{dilate}(S_k, \rho),$$

which intuitively thickens the object's boundary by ρ pixels in every direction. Two objects k and ℓ are considered adjacent if the dilated support of one intersects the original mask of the other:

$$S_k^{(\rho)} \cap S_\ell \neq \emptyset.$$

Once adjacency is established, we use the centroids $\mathbf{c}_k$ and $\mathbf{c}_\ell$ to infer the dominant orientation of ℓ with respect to k, yielding interpretable statements such as "the forceps is to the left of the needle." Importantly, this procedure complements absolute locations: when two objects are well separated across regions, their relative order is already captured (e.g., an object in the top-right is naturally to the right of one in the top-left). Pairwise relations are thus only triggered in more complex cases where objects lie within the same area and require finer geometric reasoning. Figure 2 (b) illustrates this process, showing for example how the dilated mask of the monopolar curved scissors intersects with the kidney parenchyma but not with the adjacent prograsp forceps.

Graded proximity. Whereas absolute positions provide coarse layout and pairwise adjacency captures direct contact, graded proximity introduces a softer notion of distance between instruments and tissues. For each instrument mask S_k, we generate a sequence of dilations $\{S_k^{(r_m)}\}_{m=1}^{L}$ with increasing radii r_m. For a tissue mask S_ℓ, we then record the smallest index m such that

$$S_k^{(r_m)} \cap S_\ell \neq \emptyset.$$

This index is mapped to a discrete label (e.g., "touching," "very close to," "close to," or "far from"). In this way, graded proximity enriches the spatial description by quantifying not only whether an instrument is in contact with tissue, but also the degree of closeness, offering a more nuanced account of their spatial relationship.

A.2. Interaction Heuristics

Unlike segmentation masks that require geometric reasoning, the output of the interaction model is already provided in a predefined categorical format, indicating functional relations such as "grasping," "cutting,"or "idle." The heuristics module H_{int} converts these categorical predictions into declarative textual facts using simple templates. For example:

> prograsp forceps: "Tissue Manipulation"
> monopolar curved scissors: "Idle"

This yields a fact set $\mathcal{T}_{\text{int}}$ that directly summarizes the functional roles of instruments in the current frame. Combined with the spatial cues from segmentation, these interaction statements provide complementary information, enabling the LLM to reason not only about *where* objects are but also about *how* they are used.

Appendix B. Additional Qualitative Examples

B.1. Hybrid Segmentation Illustration (SAM+FASL)

Figure 3 provides a qualitative illustration of the hybrid segmentation procedure described in Section 2.3. The region-proposal component generates a set of candidate masks, and each region is assigned a semantic class via majority voting over the predictions of the specialist segmentation model. This visualization highlights how coarse region proposals can be combined with domain-aware semantics to produce stable, coherent segmentation outputs.

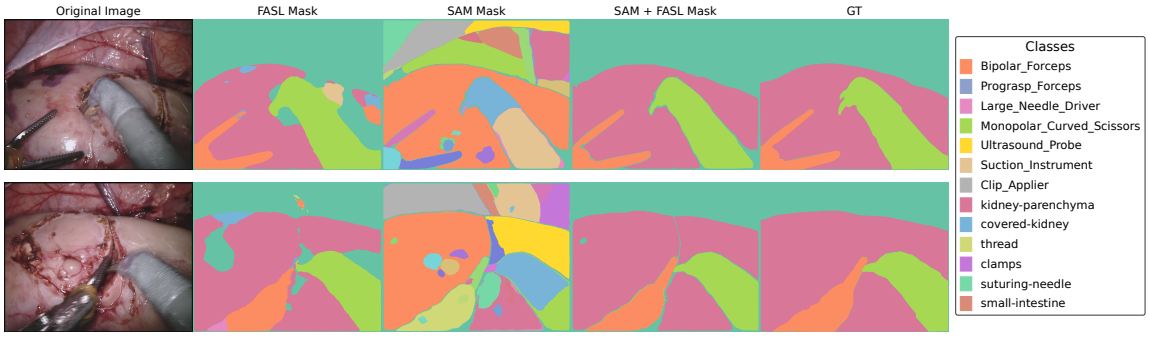

Figure 3: Example of SAM+FASL segmentation. SAM provides mask proposals, while FASL supplies domain-specific semantics. GT denotes ground truth.