

SurgXBench: Explainable Vision-Language Model Benchmark for Surgery

Jiajun Cheng¹ Xianwu Zhao¹ Sainan Liu² Xiaofan Yu¹ Ravi Prakash³
 Patrick J. Codd³ Jonathan Elliott Katz⁴ Shan Lin¹
¹Arizona State University, ²Intel Labs,
³Duke University, ⁴University of Miami

Abstract

Innovations in digital intelligence are transforming robotic surgery through more informed decision-making. Real-time awareness of surgical instrument presence and actions (e.g., cutting tissue) is essential, yet despite decades of research, most machine learning models rely on small datasets and still struggle to generalize. Recently, Vision-Language Models (VLMs) have achieved transformative advances in multimodal reasoning, suggesting strong potential for intelligent robotic surgery. However, surgical VLMs remain underexplored, and existing models show limited performance, underscoring the need for systematic benchmarks to assess their capabilities, limitations, and future development. To this end, we benchmark the zero-shot performance of several advanced VLMs on two public robotic-assisted laparoscopic datasets for instrument and action classification. Beyond standard evaluation, we integrate explainable AI to visualize VLM attention and uncover causal explanations behind predictions, providing a previously underexplored perspective for assessing model reliability. We also propose explainability-based metrics to complement standard evaluations. Our analysis reveals that surgical VLMs, despite domain-specific training, often rely on weak contextual cues rather than clinically meaningful visual evidence, highlighting the need for stronger visual and reasoning supervision in surgical applications. The code is provided in our public repository at: <https://github.com/jiajun344/SurgXBench-Explainable-Vision-Language-Model-Benchmark-for-Surgery>.

1. Introduction

Tremendous progress in robotic surgical systems has enabled high-precision, dexterous operations in minimally invasive settings, allowing access to narrow

surgical spaces that are often unreachable via traditional open surgery [3, 7]. As robotic procedures expand, integrating digital intelligence can augment surgeons' capabilities through context-aware feedback and real-time decision support [10, 36]. This integration enables more efficient procedures and more consistent, positive patient outcomes. Despite strong motivation, and over two decades of research into developing AI technologies for critical tasks such as surgical instrument, tissue, and phase recognition, existing models are still lack accuracy and robustness [3, 11, 18, 19, 21, 27, 31]. Thus, Current surgical robots remain fully controlled by human surgeons. This is largely due to the fact that most existing models are developed using small, manually annotated datasets and still struggle to consistently achieve good performance across visually diverse surgical data.

Recently, Vision-Language Models (VLMs), such as CLIP [24] and LLaVA [20], have been trained by aligning visual and textual features rather than relying on curated labels. By learning from raw image-text pairs at scale, these models achieve strong generalization across many domains [42]. In surgical robotics, early studies have developed surgical VLMs using publicly available surgical lecture videos and educational materials, but their performance remains limited and not well understood [25, 37–39]. This indicates the need for systematic benchmark studies to analyze how existing VLMs, both general and surgical models, perform on surgical tasks. A more comprehensive understanding of their strengths and limitations is essential to guide the future development of surgical VLMs.

To the best of our knowledge, the only existing VLM benchmark in surgery is by by Rau *et al.* [25], which evaluates the zero-shot and few-shot performance of several state-of-the-art VLMs on multiple surgical vision tasks, including instrument and anatomy identification, action and phase recognition, and skill assessment. Their results also show that current surgical VLMs still perform

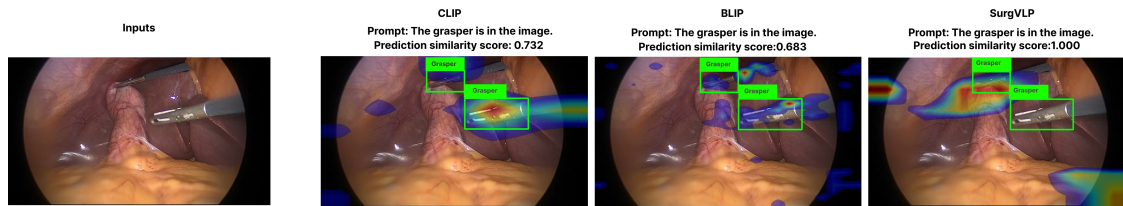


Figure 1. We evaluate two general VLMs, CLIP-ViT-B/32 [24], BLIP [17], and a surgical VLM, SurgVLP [39], on surgical instrument classification. The first image shows an input frame from the Cholec80BBox [2, 14] dataset. The next three images overlay Grad-CAM attention maps [6, 28] on top of the original images, showing each model’s focus in response to the prompt `The grasper is in the image.`. The similarity scores above each image indicate the model’s prediction confidence (min-max normalized). Green bounding boxes and labels are ground truth instrument locations and types.

poorly, sometimes even underperforming general VLMs on certain surgical tasks. Moreover, the performance of both general and surgical VLMs remains far below clinical requirements; for example, Rau *et al.* report precision below 60% for instrument recognition in both zero-shot and 5-shot settings. While Rau *et al.*’s benchmark covers many tasks, it only reports standard evaluation metrics, which primarily reflect the average poor performance across datasets, but it lacks deeper insight into why these models fail and which aspects require improvement to better adapt VLMs for surgical applications in the future.

To this end, our work goes beyond standard performance reporting. We integrate widely used explainable AI (XAI) techniques [40], including Grad-CAM [6, 28] and CLEANN [26, 29], into the evaluation of VLMs to visualize their attention and causal reasoning behind their predictions. Then, whether these models’ reasoning processes align with clinical relevant cues, such as the instrument locations and the tissues they interact with, is analyzed. Figure 1 shows a representative example from our explainability analysis: the surgical model SurgVLP [39] correctly classifies the instrument as a grasper with higher confidence (*i.e.*, higher similarity score), but compared to general VLMs such as CLIP [24] and BLIP [17], the attention of SurgVLP does not align well with the most relevant regions, which are the instrument regions in this example. This raises concerns about the reliability of the surgical VLM’s predictions, which are not adequately reflected by standard evaluation metrics. To systematically assess these observations, we further propose several evaluation metrics to quantitatively assess how well the models’ attention and decision-making processes align with key clinical cues.

As the first step toward explainability-based benchmarking for surgical tasks, we focus on two fundamental and critical tasks, surgical instrument and action classification. They serve as the foundation for many high-level applications such as surgical phase recognition and skill

assessment. Moreover, our explainability-driven analysis pipeline and proposed evaluation metrics can be flexibly extended to other surgical tasks, as long as relevant clinical cues are defined with appropriate domain knowledge. In summary, our main contributions are as follows:

- We introduce SurgXBench, the first VLM benchmarking framework for surgical tasks that integrates explainability. By analyzing model attention and causal reasoning, SurgXBench reveals a significant mismatch between where surgical VLMs focus and clinically relevant regions, even for correct predictions. This indicates the need for stronger visual grounding and semantic supervision in future models.
- We propose several new metrics to quantitatively assess attention alignment. These metrics extend our qualitative findings based on the explainability analysis and complement standard task metrics, enabling a more comprehensive evaluation of model prediction reliability.
- For instrument action classification, we leverage an off-the-shelf optical flow method with camera motion correction to automatically extract instrument-tissue interaction regions. This augments existing datasets with ground truth for where critical instrument interactions occur and enables quantitative evaluation of model attention for surgical actions, an aspect not rigorously studied before.

2. Related Work

General and Surgical VLMs: VLMs are powerful frameworks for multimodal understanding, broadly categorized into contrastive models and Large Vision–Language Models (LVLMs). Contrastive models such as CLIP [24] and BLIP [17] align visual and textual features, while LVLMs like LLaVA [20] couple vision encoders with large language models to generate text with richer reasoning. General VLMs generalize well across domains of AI, vision, and robotics by training on massive

online multimodal data, but lack the domain-specific knowledge needed for surgical environments. To address this gap, several surgical VLMs have been developed: SurgVLP [39] pioneered large-scale surgical pretraining on video lectures with CLIP-style contrastive learning; HecVL [37] introduced hierarchical supervision by pairing videos with free-form text and structured phase annotations; and PeskaVLP [38] incorporated LLM-augmented surgical knowledge with procedure-aware loss for better cross-modal temporal alignment. Yet, these models still show limited gains, often only marginally outperforming general VLMs or even underperforming, as confirmed by Rau *et al.* [25]. Since Rau’s study focused only on standard task metrics, we propose complementing such benchmarks with explainable AI to analyze the reasoning behind VLM predictions.

Explainable AI (XAI): While deep learning models achieve superior performance, their lack of interpretability poses challenges. Visualization methods address this by highlighting influential image regions and revealing evidence behind predictions. Grad-CAM [28] computes gradients w.r.t. feature maps, with extensions such as attention-based methods [5], Grad-CAM++ [4] for multiple objects, and Score-CAM [34] using output scores. Causal interpretation methods like CLEANN [26, 29] learn causal structures from attention, testing conditional independence between token pairs to identify tokens that affect outputs if masked. However, few medical XAI studies focus on surgery, and most lack explainability evaluation. Zhang *et al.* [44] emphasized the absence of unified XAI evaluation methods and standardized metrics for trustworthy AI in surgical tasks such as phase and instrument recognition, underscoring the need for clinically meaningful explainability frameworks.

3. Proposed Methods

In this section, we introduce the benchmarked surgical tasks in Section 3.1, explain the involved explainable AI [40] methods in Section 3.2, outline the questions guiding our explainability analysis in Section 3.3, and present the proposed explainability-based metrics in Section 3.4.

3.1. Task Definition

Our benchmark focuses on two fundamental and important tasks for intelligent robotic surgery: (i) *instrument classification*, and (ii) *triplet recognition*.

Instrument Classification: The goal of this task is to determine which surgical instruments are present in a given image. Awareness of the presence of instruments will enhance the control and use of them [3]. Moreover, since instrument usage is closely related to surgical

phases, accurate instrument classification also supports downstream applications such as phase recognition and skill assessment [31].

Triplet Recognition: In addition to instrument classification, triplet recognition is a task aims to also predict the associated instrument action as subject-verb-object triplets [22], such as (*scissors*, *cut*, *tissue*) triplet, describing which instrument performs what action on which target. This usage information can support downstream needs such as surgical workflow analysis and identification of potential procedure risks.

3.2. XAI Visualization Techniques

We employ three XAI methods to identify important regions for predictions. For contrastive models with ResNet backbone [13], Grad-CAM computes gradients of similarity scores with respect to features to produce spatial attention heatmaps. For LVLMs, Grad-CAM computes gradients of token logits with respect to features [43]. For transformer architectures [9], we use gradient-weighted attention rollout [5], which leverages self-attention matrices to identify contributing patches through relevance propagation. Additionally, CLEANN [26, 29] analyzes causal relationships between token pairs, discovering how visual attention causally influences language generation. Mathematical formulations are in the supplementary section A.

3.3. Problem Definition

A main objective of our work is to evaluate VLMs from a novel perspective using XAI technologies to assess whether the models’ prediction processes align with clinically relevant cues. Specifically, our explainability-based analysis focuses on the following questions:

- **(A) When the model makes a correct prediction:**
 - (i) Does its attention or causal reasoning behind its prediction focus on clinically relevant visual cues?
 - (ii) Are there cases where the model predicts correctly without attending to the relevant image regions, suggesting that the prediction may be unreliable or coincidental?
- **(B) When the model makes an incorrect prediction:**
 - (i) Is the incorrect prediction due to the model attending to irrelevant visual cues?
 - (ii) Does the model focus on relevant cues, but unable to recognize the instrument type due to lack of domain knowledge?

3.4. Explainability-based Metrics

To quantify the explainability-based analysis, as inspired by [28] and [6], we propose several metrics to measure the alignment between the visual cues utilized by the models

and clinically relevant cues, thereby helping answer questions outlined in Section 3.3. These explainability-based metrics are reported alongside standard evaluation metrics for involved tasks to provide a more comprehensive evaluation of the model. Detailed definitions for standard metrics are provided in supplementary material section D.

3.4.1. Metrics for Instrument Classification: AA, AC

Let I denote an input image of size $H \times W$. Grad-CAM is used to produce a heatmap $L \in \mathbb{R}^{H \times W}$, in which each pixel reflects the attention value of that location for the model's prediction. We then apply a percentile-based threshold τ to retain information about the relative intensity of attention across spatial locations. We pick top τ percent of attention value in the heatmap L . The predicted attention region is then defined as:

$$\mathcal{A}_\tau = \{(i,j) \mid L(i,j) \geq \text{Quantile}_{1-\tau}(L)\}. \quad (1)$$

where $L(i,j)$ denotes the pixel at i -th row and j -th column in L .

Let $\mathcal{C} = \{c_1, \dots, c_K\}$ denote the set of ground-truth class labels (e.g., surgical instruments) present in the image, and N_k is the number of appearances of class c_k in the image. For each class $c_k \in \mathcal{C}$, let $\mathcal{G}_{c_k}^{(r)} \subseteq \{1, \dots, H\} \times \{1, \dots, W\}$ denote the set of pixels in the r -th annotated region of class c_k , for $r = 1, \dots, N_k$. Let $\hat{c}$ be the class predicted by the model. We define the set of all annotated pixels (regardless of class) as:

$$\mathcal{G}_{\text{all}} = \bigcup_{k=1}^K \bigcup_{r=1}^{N_k} \mathcal{G}_{c_k}^{(r)} \subseteq \{1, \dots, H\} \times \{1, \dots, W\}.$$

We define the set of pixels corresponding to the predicted class $\hat{c}$ as:

$$\mathcal{G}_{\hat{c}} = \begin{cases} \bigcup_{c_k = \hat{c}} \bigcup_{r=1}^{N_k} \mathcal{G}_{c_k}^{(r)}, & \text{if } \hat{c} \in \mathcal{C} \\ \emptyset, & \text{otherwise} \end{cases}$$

Thresholded Attention Coverage (AC): This metric measures how much of the model's attention region overlaps with any annotated object, regardless of class correctness:

$$\text{AC} = \frac{|\mathcal{A}_\tau \cap \mathcal{G}_{\text{all}}|}{|\mathcal{A}_\tau|}. \quad (2)$$

Thresholded Attention Alignment (AA): This metric evaluates whether the model's attention aligns specifically with the annotated regions of the predicted class $\hat{c}$. If $\hat{c} \notin \mathcal{C}$, the score is defined to be zero:

$$\text{AA} = \begin{cases} \frac{|\mathcal{A}_\tau \cap \mathcal{G}_{\hat{c}}|}{|\mathcal{A}_\tau|}, & \text{if } \hat{c} \in \mathcal{C} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

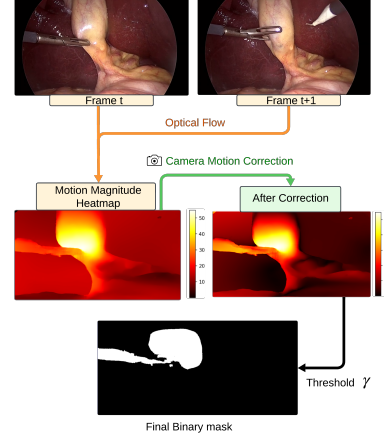


Figure 2. Pipeline of extracting instrument-tissue interaction region for benchmark triplet recognition task. Optical flow is estimated by RAFT [30], and heatmaps represent motion magnitude computed from the flow vectors. Camera correction is applied to reduce background motion in the heatmap. Threshold γ retains high motion magnitude regions to generate binary masks.

3.4.2. Metrics for Triplet Recognition: RAS

When it comes to localizing action spatial regions, things get more complex. While actions can be learned by looking at instruments [32], real applications involve tool-tissue interactions where deformable tissues carry meaningful information. Motion reflects rich regions of interest across consecutive frames—information not captured by standard semantic segmentation datasets.

RAFT and Camera Motion Correction: To approximate interaction labels, we applied RAFT [30] to obtain optical flow and identify high-motion areas. Global camera motion (Pan/Tilt/Zoom/Roll [1]) was estimated from bidirectional RAFT flow and removed via least-squares fitting. Thresholding the corrected flow with percentile γ produces binary interaction masks (Fig. 2).

RAFT Attention Score (RAS): Based on this motion-aware supervision, where $\mathcal{A}_{\tau^v}$ denotes thresholded attention heatmap of a triplet prediction, and $\mathcal{M}_{\text{raft}} \in \{0, 1\}^{H \times W}$ denotes the binary mask generated by RAFT method, $\mathcal{G}_s$ the predicted instrument pixel set, and $\mathcal{G}_{\text{all}}$ the union of all ground-truth instrument regions.

Local RAS: Measures alignment of verb attention with motion or region of the predicted instrument:

$$\text{RAS}_{\text{local}} = \frac{|\mathcal{A}_{\tau^v} \cap (\mathcal{M}_{\text{raft}} \cup \mathcal{G}_s)|}{|\mathcal{A}_{\tau^v}|}. \quad (4)$$

Model	Type	Training Data
SurgVLP	Contrastive	Surgical Video
HecVLP	Contrastive	Surgical Video
PeskaVLP	Contrastive	Surgical Video
CLIP	Contrastive	Web Image-Text
BLIP	Contrastive	Web Image-Text
LLaVA-v1.5-7B [‡]	LVL	Web Image-Text

Table 1. Models used in this study. LLaVA-v1.5-7B[‡] uses LLaMA2 as LLM backbone.

Global RAS: Measures alignment with all ground-truth instrument regions:

$$\text{RAS}_{\text{global}} = \frac{|\mathcal{A}_{\tau^v} \cap (\mathcal{M}_{\text{raft}} \cup \mathcal{G}_{\text{all}})|}{|\mathcal{A}_{\tau^v}|}. \quad (5)$$

This score quantifies how much attention of triplet predictions overlap with instrument-tissue interaction regions.

4. Experiments Set up

4.1. Dataset

We use the following two publicly available robotic-assisted laparoscopic datasets.

Cholec80BBox [2, 14] includes 5 videos with 51.7K frames annotated with instrument bounding boxes and class labels. We use this dataset to benchmark instrument classification, with class labels as ground truth for instrument classification evaluation and bounding boxes as ground truth for computing explainability-based metrics.

CholecT45 [22] consists of 45 videos with 90.5K frames annotated with frame-level triplet annotations. This dataset is used to benchmark triplet recognition. Since our $\text{RAS}_{\text{local}}$ metric (4) requires instrument bounding box annotations to measure attention alignment, we use only 5,957 frames from videos #42 and #43, which are the only ones with such annotations.

4.2. Benchmarked Models

We evaluate the zero-shot performance of six pre-trained VLMs, including three **General VLMs** CLIP, BLIP, LLaVA, along with three **Surgical VLMs** SurgVLP, HecVL, and PeskaVLP. More details of the benchmarked models are provided in Table 1.

4.3. Evaluation Strategy

We report standard classification metrics (precision, recall, F1) [23] and our attention alignment scores to assess prediction trustworthiness beyond accuracy. To address question (A-ii) from Section 3.3, we design **Attention-Based Occlusion Analysis** inspired by counterfactual explanations [12]. This also helps reveal whether VLMs

rely on shortcuts from background patterns rather than relevant regions. In addition, Grad-CAM may add noise since autoregressive models condition on previous text and visual tokens. Thus, we apply **Causal Graph Analysis** to LLM-VLMs to assess whether models use appropriate tokens for the correct task. Contrastive models compute image-prompt similarity; prompts exceeding threshold δ form the predicted set $\hat{\mathcal{P}}$, and heatmaps are generated using post-hoc XAI methods (Section 3.2) for each predicted class $\hat{p} \in \hat{\mathcal{P}}$. Implementation details are provided in Supplementary Section C.

4.3.1. Instrument Classification Task

For each image, Instrument predictions are categorized as True Positive (TP) or False Positive (FP), and we compute alignment scores using post-hoc explainability methods (Section 3.2). Let $\mathcal{P}$ denote the ground truth set for an image. For TP predictions ($\hat{p} \in \mathcal{P}$), both explainability-based metrics for instrument classification task, AA (3) and AC (2) scores are computed, while for FP predictions ($\hat{p} \notin \mathcal{P}$), only AC (2) is calculated. We report instrument-wise F1-scores on the **Cholec80BBox** dataset, along with True Positive AA and AC scores and False Positive AC scores. This enables statistical analysis of attention alignment patterns, addressing questions A-i, B-i, and B-ii from Section 3.3.

4.3.2. Triplet Recognition Task

We apply a percentile threshold γ to RAFT motion heatmaps to obtain binary masks and report attention alignment via RAS score box plots on videos 42 and 43 from the **CholecT45** dataset. We evaluate each ground truth by combining standard triplet metrics [22] with our explainability-based metrics as follows:

- **IVT match:** All components (instrument, verb, target) match in a top- k prediction. We compute $\text{RAS}_{\text{local}}$.
- **IV match:** If no IVT match, but a top- k prediction matches instrument and verb, we compute $\text{RAS}_{\text{local}}$.
- **IT match:** If no IV or IVT match, but a top- k prediction matches instrument and target, we compute $\text{RAS}_{\text{local}}$.
- **Missed:** If no top- k prediction matches the instrument, we compute average $\text{RAS}_{\text{global}}$ over all predictions.

This captures the relationship between prediction correctness and attention behavior, addressing A-i, B-i, and B-ii in Section 3.3. For LLaVA, we apply the same evaluation using all generated prediction set.

Attention-Based Occlusion Analysis: To evaluate question A-ii, we use VID42 and VID43 from CholecT45, selecting top-5 prompts as predictions. For frames with IVT-covered triplets, we apply Grad-CAM to generate heatmaps, retain top $1 - \tau$ high-attention regions using percentile threshold τ , then occlude these regions with

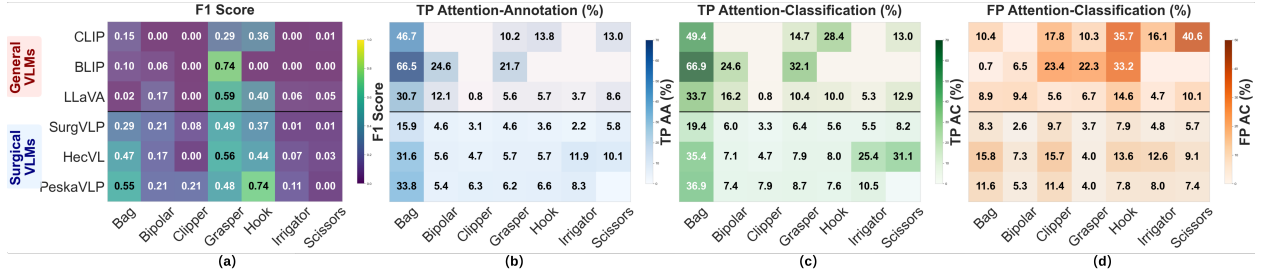


Figure 3. Instrument classification results of General VLMs (CLIP[24], BLIP[17], LLaVA[20]) and Surgical VLMs (SurgVLP[39], HecVL[37], PeskaVLP[38]) on Cholec80BBox dataset. Results are reported using standard metric (a) F1 scores, and proposed explainability-based metrics that measure model attention alignment with clinically relevant cues: (b) TP_AA, (c) TP_AC, and (d) FP_AC.

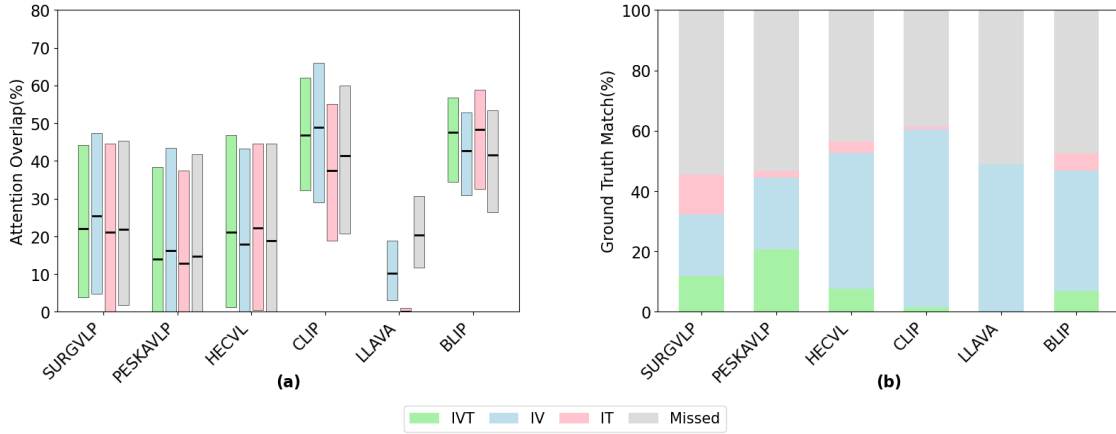


Figure 4. Triplet recognition results on CholecT45[22] dataset. (a) Box plots of RAS_{local} for IVT, IV, and IT matches, and RAS_{global} for missed cases. (b) Percentage breakdown of how ground truth triplets are matched by model predictions: IVT (complete match), IV (instrument-verb match), IT (instrument-target match), and Missed (no instrument match).

black pixels following [41]. We report effects on IVT coverage rates and similarity scores.

Causal Graph Analysis: To visualize how the model attributes predictions to visual input, we apply CLEANN to generate causal graphs from the top five attention heads which is a parallel mechanisms where each head in the transformer attention learns to focus on different visual features and relationships that attend most strongly to visual tokens during the generation of each prediction token.

4.3.3. Agreement Analysis of Clinical Evaluations and XAI-based Metrics

To assess alignment between our explainability metrics and clinical judgment, we worked with clinical partners to perform pairwise selections of Grad-CAM heatmaps that best matched surgical reasoning on 100 uniformly sampled image pairs. Pairwise comparison was chosen as it is easier for humans to compare two maps than to rank many. We repeated the same comparisons using

our metrics (AA/AC/RAS) and a baseline that selects the image with higher model confidence in the correct class. Agreement between clinician choices and each method was then quantified with Cohen’s κ [8] (higher is better).

5. Results

5.1. Instruments Attention Quality Analysis

Figure 3 shows instrument classification results evaluated using both standard and proposed explainability-based metrics. General contrastive learning VLMs outperform surgical VLMs in attention alignment with higher AA and AC scores across most instruments regardless of prediction correctness, with BLIP showing strongest attention alignment (*bag* TP_AC: 66.9%, *grasper* TP_AC: 32.1%). This suggests general contrastive learning VLMs can localize instrument regions but misclassify them due to lacking surgical knowledge, while surgical VLMs exhibit better classification performance on most

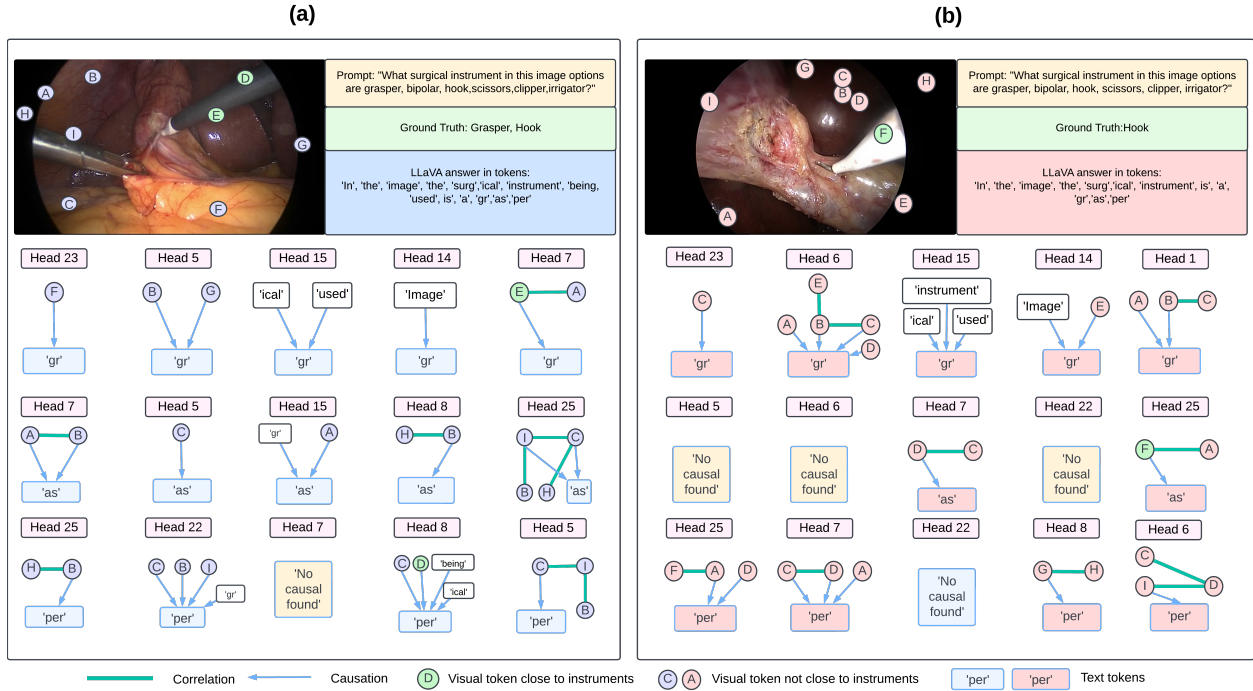


Figure 5. LLaVA predictions with CLEANN-generated causal graphs (radius 3, top-5 attention heads). (a): Correct *grasper* identification; (b): *hook* misclassified as *grasper*.

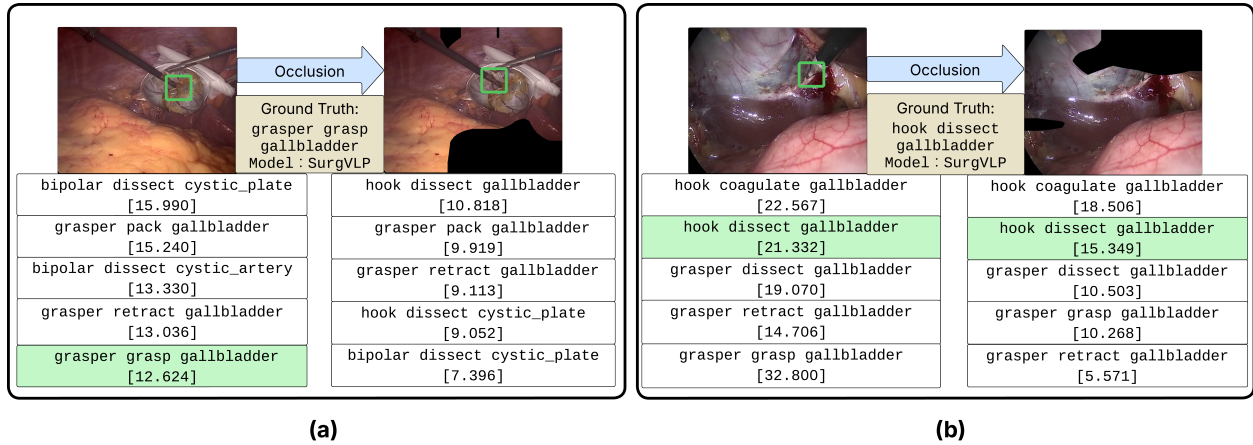


Figure 6. Counterfactual occlusion analysis results on SurgVLP. We picked two frames with the two most common instruments in the CholecT45 dataset as examples. Prompts with top-5 predicted similarity scores are shown, with the instrument tip marked by a green box in the image and corresponding prompt marked with a green background indicating IVT coverage. (a) An example of occluding irrelevant areas reducing the similarity score of the correct prediction, removing it from top-5 and leading to a wrong prediction. (b) An example of occluding relevant regions, in contrast, does not affect the prediction with correct prediction still has a relatively high similarity score.

of instruments, with PeskaVLP achieves highest F1 for multiple classes (*hook*: 0.7410, *bipolar*: 0.7899, *irrigator*: 0.7962) despite lower attention scores (*hook* TP_AA: 6.6%, TP_AC: 7.6%). This suggests that surgical VLMs may rely on visual shortcuts rather than relevant cues, we

further investigate through occlusion studies in Section 5.4. Additional detailed boxplots and precision recall results are provided in supplementary material section E.1. Additionally, the agreement analysis shows substantial agreement for AA ($\kappa = 0.651$) for AA and moderate

for AC ($\kappa = 0.579$), both higher than the conventional baseline ($\kappa = 0.489$), indicating that AA and AC better capture clinically meaningful reasoning.

5.2. Triplets Attention Quality Analysis

Figure 4 shows triplet recognition results. In Figure 4(a), contrastive general VLMs exhibit higher RAS_{local} and RAS_{global} scores (BLIP: 47.5%) than surgical VLMs (SurgVLP: 22.1%), while LLaVA shows weaker attention alignment. However, Figure 4(b) shows general VLMs struggle with tissue recognition. CLIP achieves 59.1% IV-matching but only 1.4% IVT-matching, while LLaVA produces no IVT-matched triplets (0.0%) despite 48.9% IV-matching. Surgical VLMs demonstrate stronger tissue recognition but lower overall RAS scores. PeskaVLP leads with 20.8% IVT-matching, followed by SurgVLP (11.9%) and HecVL (7.6%). Similar to findings in instrument classification, the agreement analysis shows substantial agreement for RAS ($\kappa = 0.601$) compared to baseline ($\kappa = 0.399$).

5.3. Causal Graph Analysis on LVLM

Two examples of LLaVA’s causal graph analysis on instrument classification are presented in Figure 5. Even when selecting the top five heads most focused on visual tokens, the CLEANN algorithm reveals that many text tokens still emerge as direct causal contributors to the prediction. This suggests that certain keywords, such as *surgical*, *image*, and *instrument*, may exert a direct influence to the final prediction than visual tokens. Additionally, most visual tokens do not lie within clinically relevant regions for instrument classification. Instead, they are often scattered, with no consistent spatial pattern observed across examples. We further quantify text–visual ratios in the causal graph and provide additional quantitative examples in Supplementary E.3–E.4.

5.4. Attention-based Occlusion Analysis

Figure 6 shows SurgVLP results: after occluding regions guided by attention heatmaps, 28.2% of frames with IVT (instrument–verb–target) matches lost all matches in top-5 predictions, while the rest showed a 9.87 average similarity drop for IVT prompts. Figure 6(a) illustrates matches lost despite occluding irrelevant regions, and (b) shows persistence even when instruments were fully blocked. Additional examples are in Supplementary E.2.

6. Discussion

Surgical VLM requires stronger visual supervision. General VLMs can roughly localize instruments but often fail to recognize their types. Surgical VLMs trained on

surgical video data achieve moderately higher classification accuracy, but this does not stem from better attention to clinically relevant cues; in fact, they often show worse alignment between attention and meaningful regions in our benchmarked tasks. For general VLMs, low accuracy arises from lack of surgical knowledge. Surgical VLMs, despite domain-specific training, still fail to achieve satisfactory performance and often attend less to relevant cues. One reason is that language supervision in lecture video datasets is typically provided at the clip rather than frame level, encouraging reliance on high-level contextual patterns such as visually similar organs or tissues in the background rather than precise localization of instruments, as confirmed by our occlusion experiments in Figure 6.

LVLMs in Surgical applications may require supervision on reasoning. According to the causal graph analysis, current LVLM may rely more on text tokens to make decisions, at least on the tasks that are benchmarked in this work. For many surgical tasks, the text prompt may not provide deterministic cues, and decisions should be grounded more in visual information. Therefore, this behavior is suboptimal and suggests that model training should encourage more reliance on visual cues for decision-making.

Limitations and Future work. Our work focuses on instrument and action classification, but the explainability-based benchmark pipeline can extend to tasks such as tissue/organ identification and phase understanding. We will also explore XAI in surgical video–language models [15, 16, 33, 35] and its role in detecting model failures, though defining clinically relevant cues and extracting them as ground truth for evaluating attention and causal reasoning remain open challenges.

7. Conclusion

To bridge the gap in understanding the limitations of VLMs for surgical tasks, we present SurgXBench, the first benchmark emphasizing explainability-driven analysis in this domain. We study the attention behavior and causal mechanisms of three general and three surgical VLMs, using both qualitative attention-based occlusion analysis and quantitative explainability metrics. This benchmark provides a foundation for advancing trustworthy multimodal models in surgical applications.

References

- [1] J. Almeida, R. Minetto, T. A. Almeida, R. da S. Torres, and N. J. Leite. Robust estimation of camera motion using optical flow models. In *Proceedings of the International Symposium on Visual Computing*, pages 435–446, 2009. 4

- [2] T. Abdulkaki Alshirbaji, H. Arabian, N. A. Jalal, A. Battistel, P. D. Docherty, T. Neumuth, and K. Moeller. Cholec80-boxes: Bounding-box labeling data for surgical tools in cholecystectomy images. 2024. To be submitted. 2, 5
- [3] David Bouget, Max Allan, Danail Stoyanov, and Pierre Jannin. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Medical image analysis*, 35:633–654, 2017. 1, 3
- [4] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. 3
- [5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 397–406, 2021. 3
- [6] J. Choe, S. J. Oh, S. Lee, S. Chun, Z. Akata, and H. Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3133–3142, 2020. 2, 3
- [7] Gastone Ciuti, Robert J Webster III, Ka-Wai Kwok, and Arianna Menciassi. Robotic surgery. *Nature Reviews Bioengineering*, pages 1–14, 2025. 1
- [8] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. 6
- [9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3
- [10] Pierre E Dupont, Bradley J Nelson, Michael Goldfarb, Blake Hannaford, Arianna Menciassi, Marcia K O’Malley, Nabil Simaan, Pietro Valdastrì, and Guang-Zhong Yang. A decade retrospective of medical robotics research from 2010 to 2020. *Science robotics*, 6(60):eabi8017, 2021. 1
- [11] Luis C Garcia-Peraza-Herrera, Wenqi Li, Lucas Fidon, Caspar Gruijthuijsen, Alain Devreker, George Attilakos, Jan Deprest, Emmanuel Vander Poorten, Danail Stoyanov, Tom Vercauteren, et al. Toolnet: holistically-nested real-time segmentation of robotic surgical tools. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 5717–5722. IEEE, 2017. 1
- [12] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual visual explanations. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2376–2384, 2019. 5
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3
- [14] N. A. Jalal, T. A. Alshirbaji, P. D. Docherty, H. Arabian, T. Neumuth, and K. Möller. Surgical tool classification & localisation using attention and multi-feature fusion deep learning approach. *IFAC-PapersOnLine*, 56(2):5626–5631, 2023. 2, 5
- [15] Juseong Jin and Chang Wook Jeong. Surgical-llava: Toward surgical scenario understanding via large language and vision models. *arXiv preprint arXiv:2410.09750*, 2024. 8
- [16] Jiajie Li, Garrett Skinner, Gene Yang, Brian R Quaranto, Steven D Schwartzberg, Peter CW Kim, and Jinjun Xiong. Llava-surg: towards multimodal surgical assistant via structured surgical video learning. *arXiv preprint arXiv:2408.07981*, 2024. 8
- [17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12888–12900, 2022. 2, 6
- [18] Shan Lin, Fangbo Qin, Yangming Li, Randall A Bly, Kris S Moe, and Blake Hannaford. Lc-gan: Image-to-image translation based on generative adversarial network for endoscopic images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2914–2920. IEEE, 2020. 1
- [19] Shan Lin, Fangbo Qin, Haonan Peng, Randall A Bly, Kris S Moe, and Blake Hannaford. Multi-frame feature aggregation for real-time instrument segmentation in endoscopic video. *IEEE Robotics and Automation Letters*, 6(4):6773–6780, 2021. 1
- [20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023. 1, 2, 6
- [21] Zhen-Liang Ni, Gui-Bin Bian, Zeng-Guang Hou, Xiao-Hu Zhou, Xiao-Liang Xie, and Zhen Li. Attention-guided lightweight network for real-time segmentation of robotic surgical instruments. In *2020 IEEE international conference on robotics*

- and automation (ICRA), pages 9939–9945. IEEE, 2020. 1
- [22] C. I. Nwoye, T. Yu, C. Gonzalez, B. Seeliger, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022. 3, 5, 6
- [23] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020. 5
- [24] 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 *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 1, 2, 6
- [25] Anita Rau, Mark Endo, Josiah Aklilu, Jaewoo Heo, Khaled Saab, Alberto Paderno, Jeffrey Jopling, F Christopher Holsinger, and Serena Yeung-Levy. Systematic evaluation of large vision-language models for surgical artificial intelligence. *arXiv preprint arXiv:2504.02799*, 2025. 1, 3
- [26] R. Y. Rohekar, Y. Gurwicz, and S. Nisimov. Causal interpretation of self-attention in pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 36, pages 31450–31465, 2023. 2, 3
- [27] Tobias Roß, Annika Reinke, Peter M Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, et al. Comparative validation of multi-instance instrument segmentation in endoscopy: results of the robust-mis 2019 challenge. *Medical image analysis*, 70:101920, 2021. 1
- [28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 2, 3
- [29] G. B. M. Stan, E. Aflalo, R. Y. Rohekar, et al. Lvlm-interpret: An interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*, 2024. 2, 3
- [30] Z. Teed and J. Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419, 2020. 4
- [31] Apolline Sauzéon 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. 1, 3
- [32] Arnaud Twinanda, Emre Aksan, Raphael Sznitman, Peter Wolf, Leo Gruendlinger, Christoph Jud, and Tiago da Costa. Cholect50: A benchmark for surgical action triplet recognition. *arXiv preprint arXiv:2204.05235*, 2022. 4
- [33] Guankun Wang, Wenjin Mo, Junyi Wang, Long Bai, Kun Yuan, Ming Hu, Jinlin Wu, Junjun He, Yiming Huang, Nicolas Padoy, et al. Surgvidlm: Towards multi-grained surgical video understanding with large language model. *arXiv preprint arXiv:2506.17873*, 2025. 8
- [34] Haofan Wang, Zifan Wang, Pengfei Zhao, Yanfeng Pei, Xiaoyu Wang, Alan L. Yuille, and Wei Shen. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020. 3
- [35] Nan Xi, Jingjing Meng, and Junsong Yuan. Chain-of-look prompting for verb-centric surgical triplet recognition in endoscopic videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5007–5016, 2023. 8
- [36] Michael Yip. The robot will see you now: Foundation models are the path forward for autonomous robotic surgery. *Science Robotics*, 10(104):eadt0684, 2025. 1
- [37] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. *arXiv preprint arXiv:2405.10075*, 2024. 1, 3, 6
- [38] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 3, 6
- [39] Kun Yuan, Vinkle Srivastav, Tong Yu, Joël L. Lavanchy, Jacques Marescaux, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. *arXiv preprint arXiv:2307.15220*, 2023. 1, 2, 3, 6
- [40] É. Zablocki, H. Ben-Younes, P. Pérez, and

- M. Cord. Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision*, 130(10):2425–2452, 2022. [2](#), [3](#)
- [41] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014. [6](#)
- [42] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. [1](#)
- [43] X. Zhang, Y. Quan, C. Shen, X. Yuan, S. Yan, L. Xie, W. Wang, C. Gu, H. Tang, and J. Ye. From redundancy to relevance: Information flow in lvlms across reasoning tasks. *arXiv preprint arXiv:2406.06579*, 2024. [3](#)
- [44] Yiming Zhang, Ying Weng, and Jonathan Lund. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*, 12(2):237, 2022. [3](#)