

Supplementary Material for SurgXBench: Explainable Vision-Language Model Benchmark for Surgery

Anonymous WACV Applications Track submission

Paper ID 805

001 A. Post-hoc XAI Visualization Techniques - 002 Mathematical Formulations

003 This section provides the detailed mathematical formu-
004 lations for the post-hoc XAI visualization methods de-
005 scribed in Section 3.2 of the main paper.

006 A.1. Grad-CAM for ResNet VLM

007 For VLMs that use ResNet-style backbones [4–6], the
008 Grad-CAM formulation is defined as follows. Let $I \in$
009 $\mathbb{R}^{H' \times W' \times 3}$ be the input image, and let $A \in \mathbb{R}^{K \times H \times W}$
010 denote the output of a selected convolutional layer in the
011 image encoder, where K is the number of channels and
012 $H \times W$ is the spatial resolution. Let $A_k(i, j)$ denote
013 the activation at spatial position (i, j) in channel k . Let
014 $\mathcal{P} = \{P_1, P_2, \dots, P_N\}$ be a set contain N candidate
015 natural language prompts. The vision-language model
016 computes an embedding for the image $\phi_{\text{img}}(I)$ and for
017 each text prompt $\phi_{\text{text}}(P_k)$, and calculates a similarity
018 score S_{P_k} (typically cosine similarity) for each prompt
019 in \mathcal{P} :

$$020 S_{P_k} = \frac{\phi_{\text{img}}(I)^\top \phi_{\text{text}}(P_k)}{\|\phi_{\text{img}}(I)\| \cdot \|\phi_{\text{text}}(P_k)\|}.$$

021 The predicted prompt \hat{P} is then selected as the one
022 with the highest similarity score:

$$023 \hat{P} = \arg \max_{P_k \in \mathcal{P}} S_{P_k},$$

024 and let $S_{\hat{P}}$ denote the scalar similarity score correspond-
025 ing to the predicted prompt \hat{P} . Grad-CAM computes the
026 gradient of the score S_c with respect to the feature map
027 $A_k(i, j)$, and then applies global average pooling over
028 spatial dimension:

$$029 \alpha_k = \frac{1}{H \cdot W} \sum_{i=1}^H \sum_{j=1}^W \frac{\partial S_{\hat{P}}}{\partial A_k(i, j)}.$$

The gradient tells how much each feature map channel
 A_k contributes to the model’s prediction S_c . It does
this by averaging the gradient of predicted score with
respect to all locations (i, j) in that channel and use it act
as a weight that reflects the importance of that channel.
Based on it, to compute the Grad-CAM heatmap, we first
obtain a weighted combination of the activation maps
using channel-wise importance weights α_k , followed by
a ReLU. Finally, we apply bilinearly upsampling to resize
 L to the size of original image.

$$L_{\text{Grad-CAM}} = \text{Upsample}_{\text{bilinear}} \left(\text{ReLU} \left(\sum_{k=1}^K \alpha_k \cdot A_k \right) \right). \quad (1)$$

041 A.2. Gradient-Based Attention Rollout for 042 CLIP-ViT

043 Following the method from [1], in transformer model, we
044 leverage the self-attention matrices and their gradients
045 to identify input image patches that contribute most to
046 the similarity between an image and a given text prompt.
047 Let $\mathcal{T}^{(l)} \in \mathbb{R}^{N \times N}$ denote the self-attention matrix from
048 the l -th transformer block in the ViT encoder, where N
049 is the number of tokens, a special token added at the
050 beginning of the input, and its output representation is
051 used to summarize the entire image for classification
052 tasks. We extract $\mathcal{T}^{(l)}$ after the softmax operation during
053 the forward pass.

We compute the gradient of the similarity score with
respect to the attention map:

$$056 \nabla \mathcal{T}^{(l)} = \frac{\partial S_{\hat{P}}}{\partial \mathcal{T}^{(l)}}, \quad 056$$

057 where $S_{\hat{P}}$ is the predicted similarity score corresponding
058 to the selected text prompt in a zero-shot setting. To
059 capture influence on the similarity score, we compute the

060 gradient-weighted attention map:

$$061 \quad \tilde{\mathcal{T}}^{(l)} = \text{ReLU} \left(\nabla \mathcal{T}^{(l)} \odot \mathcal{T}^{(l)} \right),$$

062 where \odot denotes the elementwise (Hadamard) product.
063 This highlights token-to-token interactions that both re-
064 ceive high attention and significantly impact the simi-
065 larity score. To propagate relevance through the trans-
066 former layers, we recursively update a relevance matrix
067 $\mathbf{R}^{(l)} \in \mathbb{R}^{N \times N}$ as follows:

$$068 \quad \mathbf{R}^{(L)} = I, \quad \mathbf{R}^{(l)} = \mathbf{R}^{(l+1)} + \tilde{\mathcal{T}}^{(l)} \mathbf{R}^{(l+1)},$$

069 starting from the final layer L , where I is the $N \times N$
070 identity matrix. This recursive formulation accumulates
071 relevance flow across all layers. Let each patch token
072 indexed by $k \in \{1, \dots, N-1\}$, after propagating to
073 the input layer, we extract the relevance scores from the
074 [CLS] token to each image patch token:

$$075 \quad r_k = \mathbf{R}_{\text{CLS}, k}^{(0)}, \quad \text{for } k = 1, \dots, N-1,$$

076 where r_k denotes the relevance of the k -th image patch
077 token to the final similarity score.

078 Finally, we reshape and upsample the vector
079 $[r_1, \dots, r_{N-1}] \in \mathbb{R}^{N-1}$ into the input image size using
080 bilinear interpolation. The final heatmap can be con-
081 structed as following:

$$082 \quad L_{\text{Grad-CAM-ViT}} = \text{Upsample}_{\text{bilinear}} \left(\text{reshape} \left([r_1, \dots, r_{N-1}] \right) \right), \quad (2)$$

083 A.3. Grad-CAM for Multimodal Large Lan- 084 guage Models

085 For large language models, we adopt a method inspired
086 by [7]. Since the predicted classes are generated as nat-
087 ural language, we apply Grad-CAM to specific tokens
088 in the output sequence. To obtain a Grad-CAM heatmap
089 for a given token t_j , we isolate its corresponding logit z_j
090 and compute the gradient with respect to the visual fea-
091 ture maps A_k from the last layer of the language model’s
092 decoder. Specifically, we compute:

$$093 \quad G_k = \frac{\partial z_j}{\partial A_k}.$$

094 The remaining steps follow the standard Grad-CAM pro-
095 cedure described earlier.

096 A.3.1. CLEANN: Causal Learning for Attention Net- 097 works

098 Unlike contrastive learning, where taking the derivative
099 of the similarity score with respect to visual tokens or

feature maps can provide straightforward visual explain- 100
ability, large vision-language models (LVLMs) involve 101
more complex reasoning and interaction processes. As 102
discussed in [7], Grad-CAM in LVLMs may indicate 103
where the model focuses during generation, but it does 104
not fully capture the underlying reasoning. As a result, 105
our attention analysis procedure for LLaVA serves more 106
as a soft interpretability measure. To better support our in- 107
vestigation, we adopt the Causal Graph Analysis method 108
CLEANN [2, 3], which provides a more principled view 109
of the model’s internal reasoning. For target token t 110
and relevant token set \mathcal{T} , the method tests conditional 111
independence between token pairs $X, Y \in \mathcal{T}$ given con- 112
ditioning set $\mathcal{Z} \subset \mathcal{T}$ using their attention distributions, 113
where \mathbf{a}_X represents the attention vector (row) of token 114
 X in the attention matrix. This approach enables the 115
discovery of how visual attention patterns causally influ- 116
ence language generation beyond simple correlation in 117
multi-modal transformer architectures. 118

119 B. Camera Motion Correction - Mathemati- 120 cal Formulation

This section provides the detailed mathematical formula- 121
tion for the camera motion correction method described in 122
the main paper. 123

124 B.1. Camera Motion Model

We model the global camera motion as a linear combi- 125
nation of four basic operations: Pan (P), Tilt (T), Zoom 126
(Z), and Roll (R). Each operation corresponds to a 127
known optical flow prototype vector at any given pixel 128
coordinate (x, y) relative to the image center: 129

$$130 \quad \text{Pan (p): } p = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \text{ Tilt (t): } t = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \text{ Zoom (z):}$$

$$131 \quad z = \begin{pmatrix} -x \\ -y \end{pmatrix}, \text{ Roll (r): } r = \begin{pmatrix} y \\ -x \end{pmatrix}.$$

The predicted flow from the camera, f_{cam} , is a 132
weighted sum of these prototypes:

$$f_{\text{cam}} = P \cdot p + T \cdot t + Z \cdot z + R \cdot r$$

132 B.2. Bidirectional Optimization

To find the actual camera motion parameters robustly, 133
we first use RAFT to compute total flow in both direc- 134
tions—the forward flow (f_{fwd}) and the backward flow 135
(f_{bwd})—and solve a least-squares problem over the valid 136
pixel domain $\Omega \subset \mathbb{R}^2$. The optimization goal is to find 137
the parameters P, T, Z, R that minimize the total squared 138
error $E = E_{\text{fwd}} + E_{\text{bwd}}$, which is the sum of the error 139
contributions from both directions. 140

141 The forward error component is:

$$142 E_{\text{fwd}} = \sum_{(x,y) \in \Omega} \|f_{\text{cam}}(x, y) - f_{\text{fwd}}(x, y)\|^2$$

$$143 = \sum_{(x,y) \in \Omega} \|(P \cdot p + T \cdot t + Z \cdot z + R \cdot r) - f_{\text{fwd}}(x, y)\|^2$$

(3)

The backward error component is:

$$E_{\text{bwd}} = \sum_{(x,y) \in \Omega} \|f_{\text{cam}}(x, y) + f_{\text{bwd}}(x, y)\|^2,$$

144 where the backward error assumes that camera motion
145 should be consistent in both temporal directions, with the
146 backward flow representing the negative of the forward
147 camera motion.

Solving this enhanced problem yields the best-fit parameters for the camera’s movement. With these parameters, we construct the pure camera flow field, f_{cam} . Finally, we correct the original forward flow by subtracting this calculated camera motion, leaving only the motion of the independent objects:

$$f_{\text{corrected}} = f_{\text{fwd}} - f_{\text{cam}}$$

148 C. Implementation Details

149 C.1. Parameters

150 All heatmaps from post-hoc XAI methods are thresh-
151 olded using percentile-based Grad-CAM, where $\tau = 0.8$
152 selects the top 20% high-attention regions. For RAFT-
153 derived binary masks, we use a motion threshold of
154 $\gamma = 0.8$. Across CLIP, BLIP, SurgVLP, HecVL, and
155 PeskaVLP, a class is predicted if its similarity score ex-
156 ceeds the 90th percentile ($\delta = 0.9$). For the triplet recog-
157 nition task, due to the large number of class combina-
158 tions, we select the top 5 classes with highest similarity
159 scores as predictions. CLEANN is implemented with
160 key parameters: attention threshold $\kappa = 0.01$ to filter
161 relevant tokens; p-value threshold $\alpha = 10^{-5}$ to enforce
162 strong statistical significance in conditional independence
163 tests; sample size $n = 128$ for reliable inference; and
164 a cap of 50 image tokens per analysis to balance effi-
165 ciency and coverage. Rather than analyzing all attention
166 heads, we select the top-5 heads that prioritize visual
167 over textual content. Head importance is computed as
168 $\text{head_importance} = \frac{\max(\text{image_attention})}{\max(\text{text_attention})}$ where higher val-
169 ues indicate greater visual focus for the target prediction.

170 C.2. Prompts Design

171 To align with VLMs’ prompt-based input format, we
172 construct template-based prompts. For instrument clas-
173 sification, we use: “What surgical tools do you see?”

Choose from: Grasper, Bipolar, Hook, Scissors, Clip-
per, Irrigator, Bag” covering all instrument classes. For
triplet reasoning, we design a detailed prompt that guides
the model to identify surgical actions by providing instru-
ment capabilities and target options. The prompt instructs
the model to look at the surgical image carefully and iden-
tify all visible actions, providing specific lists of what
each instrument can do (e.g., grasper can grasp, retract,
dissect, or manipulate tissue) and all possible surgical
targets (cystic_plate, gallbladder, omentum, etc.). We
constrain the output to a three-word format: “instrument
verb target”.

D. Metrics

D.1. Standard Instrument Classification Metrics

Let TP, FP, and FN denote the number of true positives,
false positives, and false negatives across all instrument
classes. The metrics are computed as: Precision =
TP/(TP + FP), Recall = TP/(TP + FN), $F1 = 2 \cdot$
Precision · Recall/(Precision + Recall).

D.2. Standard Triplet Recognition Metrics:

We adopt the standard triplet metrics from [?]. Each
predicted action is represented as a triplet of class la-
bels $\langle \text{instrument}, \text{verb}, \text{target} \rangle = (s, v, o)$, with the cor-
responding ground-truth triplet denoted as (s^*, v^*, o^*) .
Binary match indicators are defined as:

$$\text{Match}(s, v, o; s^*, v^*, o^*) = (\mathbb{1}_{s=s^*}, \mathbb{1}_{v=v^*}, \mathbb{1}_{o=o^*}).$$

The standard evaluation metrics focus on three specific
matching conditions: IVT (instrument–verb–target), IV
(instrument–verb), and IT (instrument–target). These
metrics provide a finer understanding of the model’s abil-
ity to recognize structured surgical actions, which are
inherently instrument-centric; the verb and target gain
full semantic meaning only in the context of the acting
instrument.

E. Additional Results

E.1. Instrument classification

Addition results include pericision, recall and box plot
of attentiona quality metrics for instrument classification
result are provided in figure 1

E.2. Occlusion Analysis

Additional results for the occlusion analysis are pro-
vided in Figure 2 showing successful tool alignment
under occluded conditions. The left panel shows cases
where occlusion blocks irrelevant areas, yet the correct

218 IVT(instrument-verb-target) match is no longer among
 219 the top-5 predictions. The right panel shows cases where
 220 relevant regions are occluded, but the correct prediction
 221 still appears in the top-5. Prompts highlighted in green in-
 222 dicate triplets that achieve IVT coverage with the ground
 223 truth, while gray prompts represent cases where the in-
 224 strument was not correctly identified. The green bound-
 225 ing box on the images represents the instrument-relevant
 226 area in the triplets.

227 E.3. Causal Graph Analysis

228 Two additional results for triplet recognition are shown
 229 in Figure 3. These figures highlight same findings dis-
 230 cussed in Section 5 of the main paper, causal graphs
 231 generated using the CLEANN method (radius 3, top-5
 232 visually important attention heads). Generated responses
 233 are in token format where each box contains the causal
 234 graph of each head for each token. Green circles indicate
 235 visual tokens that are in or close to the relevant area, and
 236 white boxes represent text tokens that were previously
 237 generated.

238 E.4. Text Token Ratio Analysis

239 Follow the same implementation details, we conducted a
 240 text token ratio analysis using a CLEANN-based causal
 241 selection approach. This method identifies tokens that
 242 directly influence target generation through conditional
 243 independence testing and computes text-visual ratios
 244 among these causal dependencies. The results reveal
 245 substantial text over-reliance: for example, tokens such
 246 as `cutting` exhibited 77.6% text dominance, while
 247 others like `c` (from “cystic”) and `sc` (from “scissors”) also
 248 showed notable text bias. These findings confirm that
 249 surgical vision-language models often depend heavily on
 250 linguistic context rather than visual evidence, even when
 251 processing visually grounded terms. Detailed results are
 252 summarized in Table 1.

Table 1. Causal Analysis (CLEANN) of Text Dominance in Token Generation

Subword Token	Text Dom.	Visual Dom.
<code>cut</code>	0.428	0.572
<code>yst</code> [†]	0.333	0.667
<code>c</code> [†]	0.690	0.310
<code>plate</code>	0.240	0.760
<code>re</code> [†]	0.495	0.505
<code>cutting</code>	0.776	0.224
<code>sc</code> [†]	0.612	0.388
<code>gr</code> [†]	0.294	0.706

[†] Subword fragment from tokenization (e.g., `yst` from “cystic”, `re` from “retract”, `gr` from “grasper”, `sc` from “scissors”)

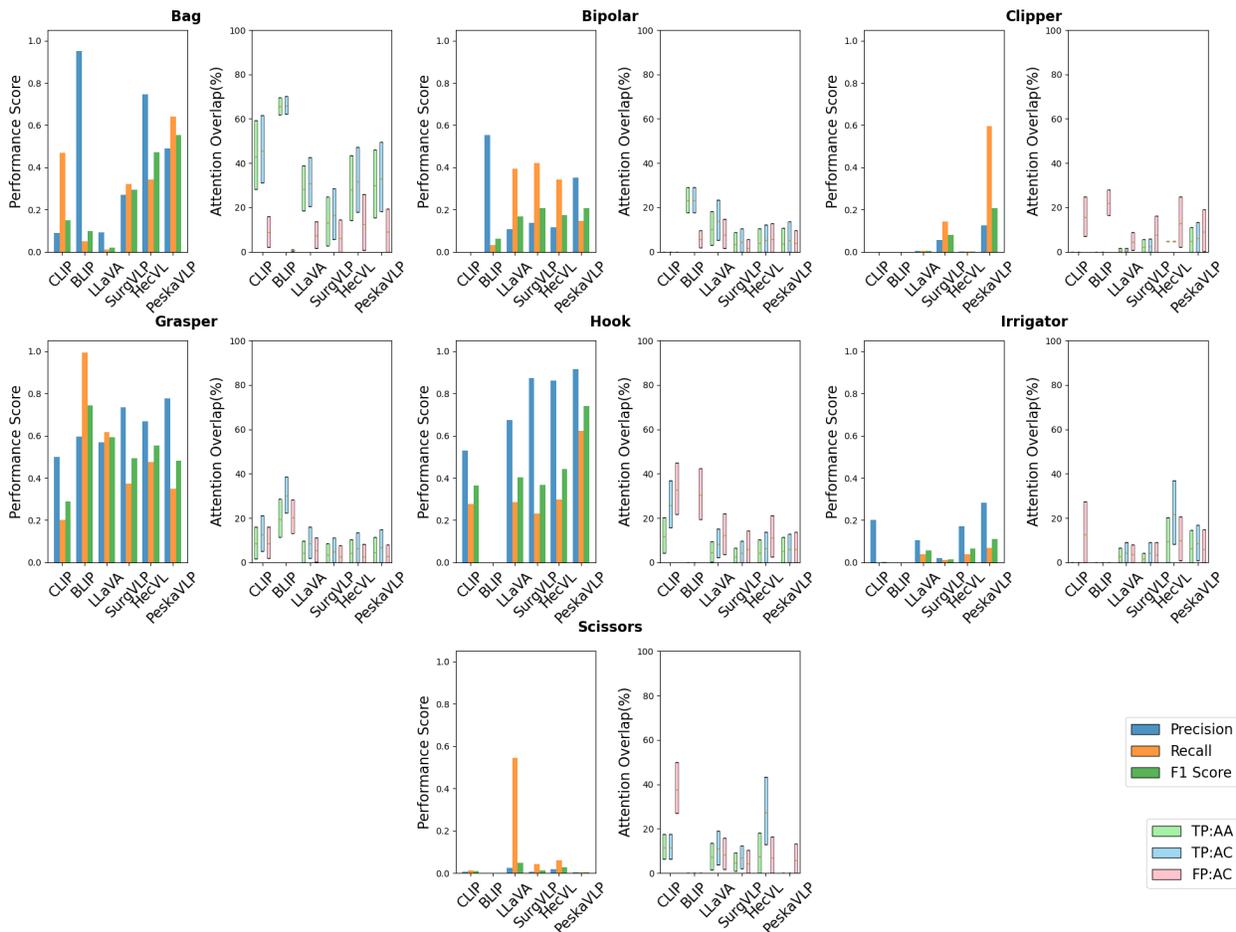


Figure 1. Surgical instrument detection performance and attention alignment across seven instrument types from the **Cholec80BBox** dataset. For each instrument, the left subplot shows classification performance (Precision, Recall, F1 Score) and the right subplot displays attention overlap AA and AC score between heatmaps and ground truth regions as box plots for True Positive and False Positive predictions.

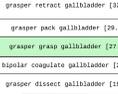
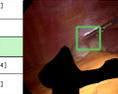
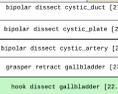
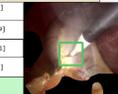
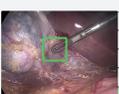
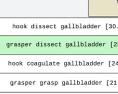
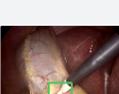
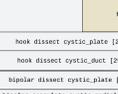
	Ground Truth: grasper grasp gallbladder Model: PeckaVLP		grasper retract gallbladder [32.841]	grasper pack gallbladder [29.540]	grasper grasp gallbladder [27.525]	bipolar coagulate gallbladder [22.454]	grasper dissect gallbladder [19.890]	Ground Truth: hook dissect gallbladder Model: PeckaVLP		grasper retract liver [20.346]	hook dissect omentum [15.486]	grasper retract omentum [14.464]	grasper pack gallbladder [8.568]	irrigator irrigate abdominal_wall_cavity [6.530]	hook dissect cystic_plate [9.381]	bipolar dissect cystic_plate [8.174]	hook dissect cystic_duct [7.683]	bipolar dissect cystic_duct [7.013]	scissors dissect gallbladder [6.990]	hook dissect cystic_plate [4.953]	hook dissect cystic_duct [3.563]	bipolar coagulate liver [2.278]	hook dissect gallbladder [1.594]	bipolar dissect cystic_duct [1.474]
	Ground Truth: hook dissect gallbladder Model: PeckaVLP		bipolar dissect cystic_duct [27.148]	bipolar dissect cystic_plate [24.548]	bipolar dissect cystic_artery [24.453]	grasper retract gallbladder [23.330]	hook dissect gallbladder [22.084]	Ground Truth: grasper retract gallbladder Model: SurgVLP		bipolar dissect cystic_duct [26.797]	hook dissect cystic_duct [23.977]	bipolar dissect cystic_artery [22.973]	bipolar dissect cystic_plate [21.041]	bipolar dissect adhesion [10.696]	hook dissect cystic_artery [40.556]	hook dissect cystic_duct [28.154]	hook dissect gallbladder [26.727]	grasper retract cystic_plate [24.580]	scissors cut cystic_artery [22.242]	hook dissect cystic_artery [36.131]	hook dissect gallbladder [34.451]	hook dissect cystic_duct [32.721]	grasper retract cystic_plate [30.711]	scissors cut cystic_artery [29.890]
	Ground Truth: grasper grasp gallbladder Model: SurgVLP		grasper grasp gallbladder [10.439]	grasper null_verb null_target [9.357]	grasper dissect gallbladder [6.100]	grasper retract gallbladder [6.125]	grasper grasp specimen_bag [4.733]	Ground Truth: grasper retract gallbladder Model: SurgVLP		grasper null_verb null_target [10.412]	grasper retract liver [10.210]	grasper grasp gallbladder [7.692]	grasper dissect gallbladder [4.700]	grasper grasp specimen_bag [4.222]	bipolar dissect gallbladder [22.387]	bipolar dissect cystic_plate [21.880]	bipolar dissect adhesion [18.666]	scissors dissect gallbladder [17.798]	bipolar dissect cystic_duct [17.276]	bipolar dissect adhesion [17.440]	bipolar dissect gallbladder [15.095]	scissors dissect gallbladder [14.506]	bipolar dissect cystic_plate [11.493]	grasper dissect gallbladder [10.674]
	Ground Truth: grasper retract gallbladder Model: SurgVLP		hook dissect cystic_duct [38.891]	hook dissect cystic_artery [35.861]	grasper retract cystic_plate [33.816]	grasper dissect gallbladder [32.101]	grasper retract gallbladder [30.639]	Ground Truth: grasper retract gallbladder Model: SurgVLP		hook dissect cystic_duct [32.772]	hook dissect cystic_artery [28.765]	grasper dissect gallbladder [27.220]	hook dissect gallbladder [22.991]	grasper retract cystic_plate [22.892]	hook dissect cystic_duct [40.946]	hook dissect cystic_artery [35.443]	scissors cut cystic_duct [31.485]	grasper retract cystic_plate [29.469]	grasper retract gallbladder [28.532]	hook dissect cystic_plate [24.824]	hook dissect cystic_artery [20.280]	grasper retract liver [17.567]	irrigator dissect cystic_pedicle [17.549]	irrigator irrigate cystic_pedicle [15.815]
	Ground Truth: grasper retract gallbladder Model: SurgVLP		hook dissect gallbladder [16.828]	hook coagulate gallbladder [15.927]	grasper dissect gallbladder [12.356]	grasper grasp gallbladder [8.231]	grasper retract gallbladder [5.384]	Ground Truth: grasper grasp gallbladder Model: SurgVLP		hook null_verb null_target [-2.992]	hook dissect gallbladder [-3.918]	irrigator dissect cystic_pedicle [-9.342]	hook coagulate gallbladder [-7.314]	irrigator irrigate cystic_pedicle [-7.583]	hook dissect cystic_duct [16.828]	hook coagulate gallbladder [15.927]	grasper dissect gallbladder [12.356]	grasper grasp gallbladder [8.231]	grasper retract gallbladder [5.384]	hook null_verb null_target [-2.992]	hook dissect gallbladder [-3.918]	irrigator dissect cystic_pedicle [-9.342]	hook coagulate gallbladder [-7.314]	irrigator irrigate cystic_pedicle [-7.583]
	Ground Truth: grasper grasp gallbladder Model: SurgVLP		grasper dissect gallbladder [21.798]	irrigator irrigate cystic_pedicle [19.078]	hook dissect gallbladder [17.944]	grasper grasp gallbladder [15.758]	hook coagulate gallbladder [13.918]	Ground Truth: grasper grasp gallbladder Model: SurgVLP		irrigator irrigate cystic_pedicle [16.254]	hook dissect gallbladder [12.932]	grasper dissect gallbladder [11.042]	irrigator dissect cystic_pedicle [9.779]	hook coagulate gallbladder [8.447]	grasper dissect gallbladder [21.798]	irrigator irrigate cystic_pedicle [19.078]	hook dissect gallbladder [17.944]	grasper grasp gallbladder [15.758]	hook coagulate gallbladder [13.918]	irrigator irrigate cystic_pedicle [16.254]	hook dissect gallbladder [12.932]	grasper dissect gallbladder [11.042]	irrigator dissect cystic_pedicle [9.779]	hook coagulate gallbladder [8.447]
	Ground Truth: grasper grasp gallbladder Model: SurgVLP		hook dissect gallbladder [30.196]	grasper dissect gallbladder [28.557]	hook coagulate gallbladder [24.890]	grasper grasp gallbladder [21.581]	clipper clip cystic_artery [18.884]	Ground Truth: grasper grasp gallbladder Model: SurgVLP		hook null_verb null_target [34.630]	grasper null_verb null_target [7.442]	grasper grasp specimen_bag [-3.609]	hook dissect gallbladder [-5.847]	hook coagulate gallbladder [-8.700]	hook dissect gallbladder [30.196]	grasper dissect gallbladder [28.557]	hook coagulate gallbladder [24.890]	grasper grasp gallbladder [21.581]	clipper clip cystic_artery [18.884]	hook null_verb null_target [34.630]	grasper null_verb null_target [7.442]	grasper grasp specimen_bag [-3.609]	hook dissect gallbladder [-5.847]	hook coagulate gallbladder [-8.700]
	Ground Truth: grasper retract gallbladder Model: SurgVLP		grasper retract gallbladder [30.218]	grasper grasp gallbladder [49.724]	grasper dissect gallbladder [49.618]	grasper pack gallbladder [43.901]	scissors dissect gallbladder [43.129]	Ground Truth: grasper retract gallbladder Model: SurgVLP		grasper retract omentum [24.439]	grasper retract gut [23.466]	grasper retract liver [20.563]	grasper grasp specimen_bag [20.370]	grasper grasp gallbladder [20.273]	grasper retract gallbladder [30.218]	grasper grasp gallbladder [49.724]	grasper dissect gallbladder [49.618]	grasper pack gallbladder [43.901]	scissors dissect gallbladder [43.129]	grasper retract omentum [24.439]	grasper retract gut [23.466]	grasper retract liver [20.563]	grasper grasp specimen_bag [20.370]	grasper grasp gallbladder [20.273]
	Ground Truth: grasper retract gallbladder Model: PeckaVLP		grasper pack gallbladder [39.254]	grasper grasp gallbladder [38.469]	grasper dissect gallbladder [37.445]	bipolar coagulate gallbladder [36.493]	grasper retract gallbladder [26.620]	Ground Truth: grasper retract gallbladder Model: PeckaVLP		grasper pack gallbladder [34.680]	grasper grasp gallbladder [34.193]	bipolar coagulate gallbladder [34.023]	grasper dissect gallbladder [32.537]	bipolar dissect gallbladder [22.244]	hook dissect cystic_duct [62.233]	grasper retract cystic_plate [58.958]	grasper retract gallbladder [58.357]	grasper grasp gallbladder [57.583]	scissors cut cystic_duct [57.490]	hook dissect cystic_duct [60.226]	scissors cut cystic_duct [57.879]	grasper retract gallbladder [56.937]	grasper retract cystic_plate [55.756]	grasper grasp gallbladder [55.620]
	Ground Truth: hook dissect gallbladder Model: PeckaVLP		hook dissect cystic_plate [29.959]	hook dissect cystic_duct [29.114]	bipolar dissect cystic_plate [28.612]	bipolar coagulate cystic_pedicle [28.002]	hook dissect gallbladder [21.351]	Ground Truth: hook dissect gallbladder Model: PeckaVLP		hook dissect cystic_plate [29.829]	bipolar dissect cystic_plate [27.144]	hook dissect cystic_duct [26.894]	bipolar coagulate liver [22.897]	bipolar dissect cystic_duct [22.432]	hook dissect cystic_duct [31.619]	hook coagulate gallbladder [30.578]	irrigator dissect cystic_pedicle [29.726]	hook dissect omentum [28.635]	hook dissect gallbladder [28.195]	hook dissect omentum [32.292]	hook dissect cystic_duct [24.920]	hook dissect gallbladder [24.795]	hook coagulate gallbladder [24.112]	bipolar coagulate liver [22.890]
	Ground Truth: grasper retract gallbladder Model: PeckaVLP		grasper grasp gallbladder [10.439]	grasper null_verb null_target [9.357]	grasper dissect gallbladder [4.150]	grasper retract gallbladder [6.125]	grasper grasp specimen_bag [4.733]	Ground Truth: grasper retract gallbladder Model: PeckaVLP		grasper null_verb null_target [10.412]	grasper retract liver [10.210]	grasper grasp gallbladder [7.692]	grasper dissect gallbladder [4.720]	grasper grasp specimen_bag [4.222]	grasper dissect gallbladder [47.723]	grasper retract gallbladder [45.302]	hook dissect cystic_duct [45.293]	hook dissect cystic_artery [43.922]	grasper retract liver [43.868]	grasper dissect gallbladder [38.200]	hook dissect omentum [37.726]	grasper retract liver [36.622]	hook dissect gallbladder [34.940]	grasper retract gallbladder [34.559]
	Ground Truth: grasper retract gallbladder Model: SurgVLP		grasper retract gallbladder [49.042]	grasper grasp gallbladder [48.278]	grasper dissect gallbladder [48.200]	grasper pack gallbladder [44.518]	scissors dissect gallbladder [43.874]	Ground Truth: grasper retract gallbladder Model: SurgVLP		grasper retract gut [22.080]	bipolar dissect adhesion [19.721]	grasper retract liver [18.200]	scissors dissect gallbladder [14.222]	hook dissect omentum [13.757]	grasper retract gallbladder [49.042]	grasper grasp gallbladder [48.278]	grasper dissect gallbladder [48.200]	grasper pack gallbladder [44.518]	scissors dissect gallbladder [43.874]	grasper retract gallbladder [22.080]	bipolar dissect adhesion [19.721]	grasper retract liver [18.200]	scissors dissect gallbladder [14.222]	hook dissect omentum [13.757]

Figure 2. Occlusion analysis results for SurgVLP, HecVLP, and PeskaVLP.

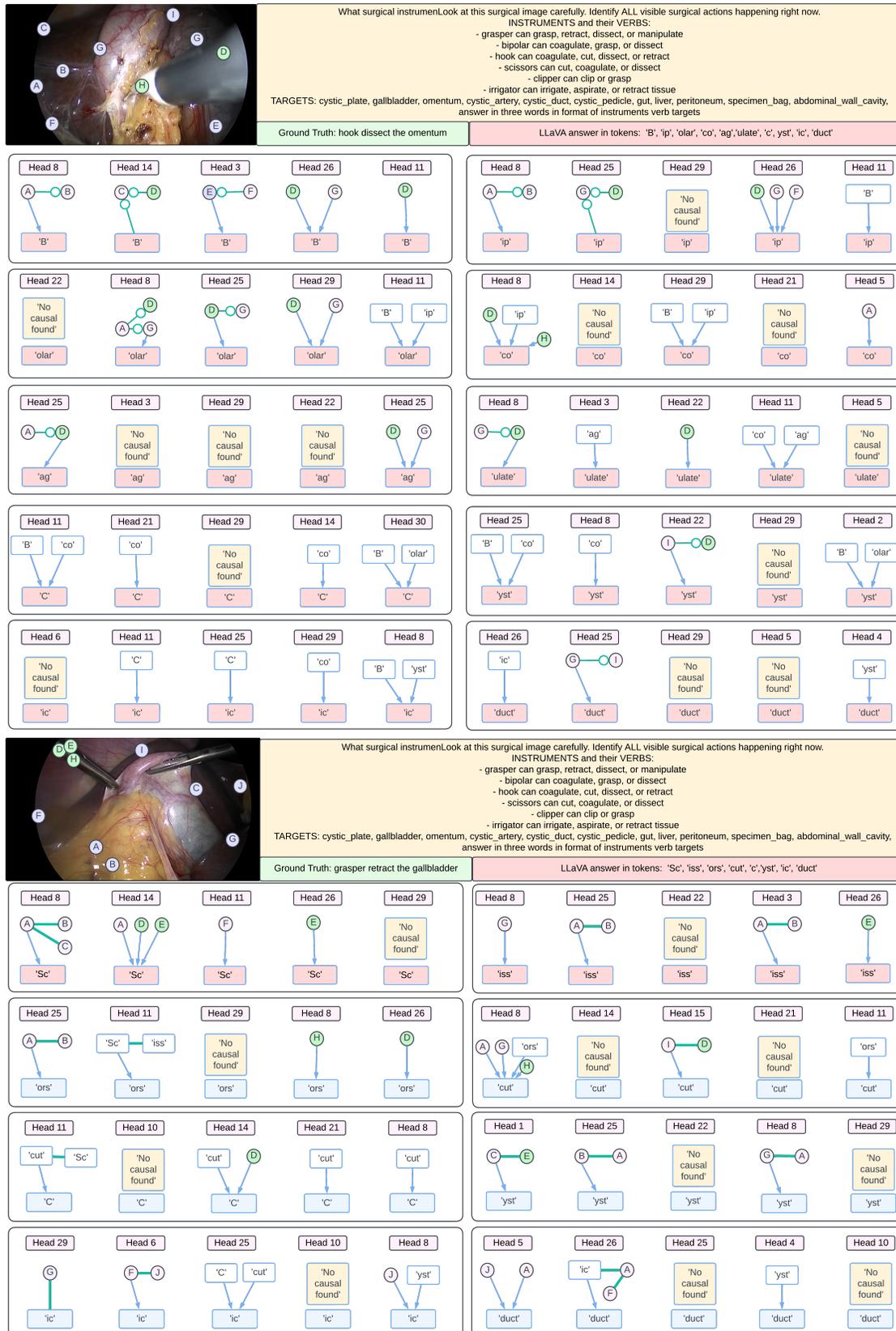


Figure 3. This figure shows triplet recognition examples from LLaVA

253 **References**

- 254 [1] Hila Chefer, Shir Gur, and Lior Wolf. Generic
255 attention-model explainability for interpreting bi-
256 modal and encoder-decoder transformers. In *Pro-
257 ceedings of the IEEE/CVF international conference
258 on computer vision*, pages 397–406, 2021.
- 259 [2] R. Y. Rohekar, Y. Gurwicz, and S. Nisimov. Causal
260 interpretation of self-attention in pre-trained trans-
261 formers. In *Advances in Neural Information Pro-
262 cessing Systems*, volume 36, pages 31450–31465,
263 2023.
- 264 [3] G. B. M. Stan, E. Aflalo, R. Y. Rohekar, et al. Lvlm-
265 interpret: An interpretability tool for large vision-
266 language models. *arXiv preprint arXiv:2404.03118*,
267 2024.
- 268 [4] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nico-
269 las Padoy. Hecvl: Hierarchical video-language pre-
270 training for zero-shot surgical phase recognition.
271 *arXiv preprint arXiv:2405.10075*, 2024.
- 272 [5] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nico-
273 las Padoy. Procedure-aware surgical video-language
274 pretraining with hierarchical knowledge augmenta-
275 tion. In *Proceedings of the 38th Conference on Neu-
276 ral Information Processing Systems (NeurIPS)*, 2024.
- 277 [6] Kun Yuan, Vinkle Srivastav, Tong Yu, Joël L. La-
278 vanchy, Jacques Marescaux, Pietro Mascagni, Nassir
279 Navab, and Nicolas Padoy. Learning multi-modal
280 representations by watching hundreds of surgical
281 video lectures. *arXiv preprint arXiv:2307.15220*,
282 2023.
- 283 [7] X. Zhang, Y. Quan, C. Shen, X. Yuan, S. Yan, L. Xie,
284 W. Wang, C. Gu, H. Tang, and J. Ye. From redun-
285 dancy to relevance: Information flow in lvlms across
286 reasoning tasks. *arXiv preprint arXiv:2406.06579*,
287 2024.