

SurgZSD: Zero-Shot Surgical Action Triplet Detection via Attribute Composition and Clinical Feasibility Inference

Zuxing Deng^{1†}, Ling Li^{1,2†(✉)}, Xiaojian Li^{1,2}, Yuting Liu¹, Xiang Ma¹, and Wenrui Zhou¹

¹ School of Management, Hefei University of Technology, Hefei 230009, China

² Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei 230009, China
cynerelee@hfut.edu.cn

Abstract. While traditional surgical action triplet detection has achieved remarkable progress, these closed-set models often fail in highly variable, real-world clinical environments where completely novel instruments, anatomical structures, and actions frequently emerge. To address these limitations, we introduce SurgZSD, a novel framework that leverages medical priors to facilitate attribute composition and joint inference for surgical triplets. Specifically, we initially fine-tune Grounding DINO to establish the precise spatial grounding of the surgical instruments. Following this localization stage, we introduce the Medical Knowledge-Guided Attribute Composition (MKAC) module. Guided by medical priors, this module leverages a pretrained Vision-Language Model (VLM) to parse the localized bounding box regions into structured semantic attributes, thereby composing unseen surgical instruments, targets, and verbs. Subsequently, to accomplish comprehensive surgical triplet recognition, we introduce the Clinical Feasibility-Constrained Joint Inference (CFCJI) module. This module introduces clinical feasibility matrices constructed based on medical priors. By formulating a Top- K joint inference objective function, it globally integrates the visual parsing scores with a clinical compatibility penalty term to yield the final surgical action triplets. Extensive experiments demonstrate that SurgZSD not only achieves state-of-the-art performance in zero-shot scenarios but also retains competitive performance under closed-set settings. The source code is available at <https://github.com/DengNova/SurgZSD>.

Keywords: Surgical Understanding · Zero-Shot Detection · Surgical Action Triplet · Attribute Composition

1 Introduction

Scene understanding in minimally invasive surgery is a fundamental prerequisite for advancing intraoperative decision support, surgical quality assessment,

[†] Equal contribution.

^(✉) Corresponding author.

and automated analysis [3, 6, 15, 20]. Distinct from coarse-grained tasks like surgical phase or instrument classification, surgical action triplet detection captures the fine-grained semantics in the form of $\langle \text{instrument}, \text{verb}, \text{target} \rangle$ [9]. By spatially grounding these interactions, it provides clinicians with a highly structured understanding, directly facilitating the targeted retrieval of specific maneuvers and the retrospective review of tissue-handling techniques [12, 18]. Nevertheless, minimally invasive surgery exhibits substantial procedural diversity: different procedures involve distinct anatomical structures and therefore often require different combinations of instruments and sequences of verbs. In real-world clinical deployments, models are inevitably confronted with novel instruments, targets, and verbs absent from the training distribution. Consequently, to ensure practical clinical utility, it is imperative that surgical action triplet detection frameworks possess zero-shot capabilities, enabling robust inference over these unseen categories [21].

Recent advancements in surgical scene understanding, typified by Surgical-VQLA [1, 2, 5] and surgical action triplet detection [13, 18], have successfully achieved simultaneous recognition and localization of instruments. Nevertheless, a critical bottleneck remains: their inability to generalize across different types of minimally invasive surgeries. Although recent efforts have begun to explore the zero-shot potential of surgical scene understanding using vision-language foundation models such as CLIP [16], existing methods typically produce only category labels and do not address spatial object localization [17]. In the general domain, open-set object detection models, exemplified by Grounding DINO [7], have demonstrated remarkable zero-shot capabilities. However, these models are typically pre-trained on large-scale, general-purpose datasets. Due to the severe domain gap and the lack of specialized training in the surgical domain, they experience significant performance degradation when directly transferred to surgical contexts [14].

Although existing models struggle to detect entities in surgical scenes when they belong to unseen classes, clinicians can often recognize previously unseen instruments or anatomical structures by leveraging medical knowledge [23]. For instance, domain knowledge indicates that bipolar forceps typically have salient visual attributes such as a yellow insulating block; these distinctive cues enable clinicians to identify bipolar forceps even without prior exposure [5]. Drawing inspiration from compositional zero-shot learning [8], we characterize surgical instruments and anatomical structures as sets of attributes, thereby establishing a foundation for zero-shot surgical action triplet detection. Distinct from generic compositional zero-shot learning frameworks, we construct a medical attribute space tailored to the surgical domain, leveraging it to enable robust inference on unseen entities, as shown in Fig. 1.

Our main contributions are as follows:

1. We propose SurgZSD, the first method to introduce zero-shot detection into the surgical domain. By leveraging medical prior knowledge, SurgZSD simplifies zero-shot recognition into a composition problem over salient attributes.

2. We propose the Medical Knowledge-Guided Attribute Composition (MKAC) module to enable the zero-shot recognition of unseen instruments, targets, and verbs. This module leverages a VLM to parse localized regions into structured attributes and match them with predefined medical prototypes, thereby composing unseen surgical entities.
3. We propose the Clinical Feasibility-Constrained Joint Inference (CFCJI) module to achieve comprehensive surgical triplet recognition. This module formulates a Top- K joint inference objective to globally integrate visual parsing scores with clinical feasibility constraints, thereby yielding the final optimal surgical triplets.

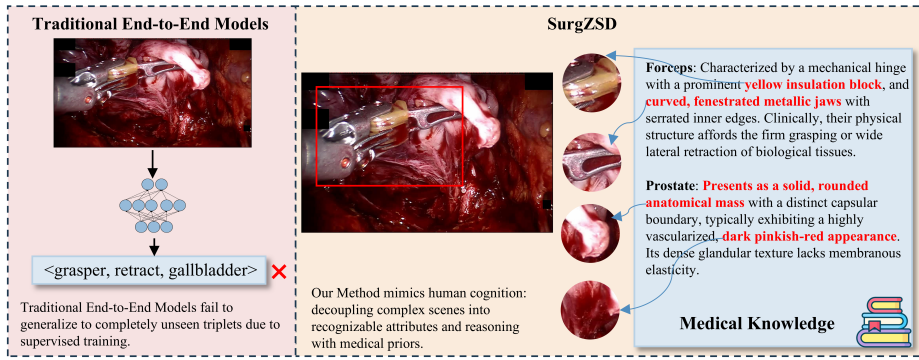


Fig. 1: **The illustration of our motivation.** The left panel demonstrates that traditional surgical action triplet detection models often output incorrect triplet predictions when confronted with zero-shot scenarios. The right panel further summarizes our core idea: we introduce medical prior knowledge to parse key visual cues of instruments and anatomical structures into attributes, and perform compositional reasoning over these attributes to enable zero-shot recognition of unseen categories.

2 Methods

2.1 Zero-shot surgical action triplet detection task definition

Current closed-set triplet detection struggles with severe domain shifts when encountering novel procedures [19]. To address this, we formalize a strict zero-shot paradigm where the training ($\mathcal{C}_s$) and testing ($\mathcal{C}_u$) classes are completely disjoint ($\mathcal{C}_s \cap \mathcal{C}_u = \emptyset$). Specifically, our testing set evaluates entirely unseen instruments (aspirator, clip applicator, forceps, needle driver) and targets (fluid, thread, bladder, fascias, prostate, seminal vesicle). Furthermore, due to the significant discrepancy in surgical procedures between the datasets, all evaluated actions are also completely unobserved during training.

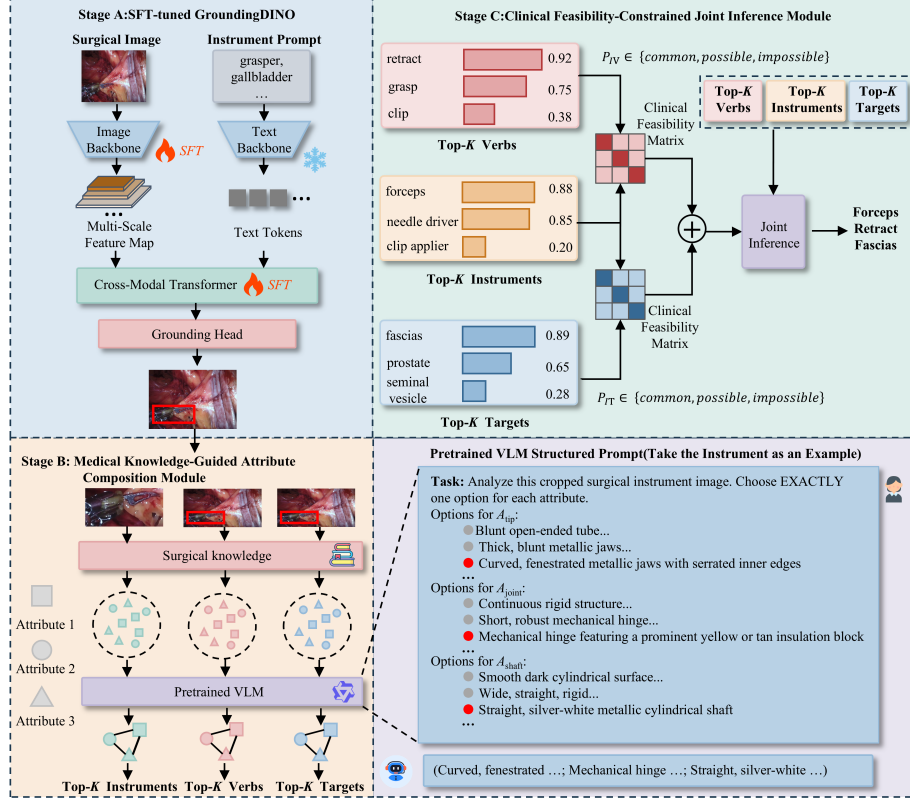


Fig. 2: Overview of the proposed SurgZSD framework. The pipeline consists of three synergistic stages. Stage A: A fine-tuned Grounding DINO accurately localizes surgical instruments. Stage B: Guided by medical knowledge, a VLM parses the attributes of the instruments within the bounding boxes to recognize unseen categories via semantic composition. Using these bounding boxes as visual prompts, the VLM further parses and composes the attributes of the interacting targets and surgical actions. The questions and options shown in the figure are illustrated using the instrument branch as an example. In the actual pipeline, the model performs analogous attribute queries for the instrument, target, and verb, respectively. Stage C: A Top-K joint inference process, strictly constrained by clinical feasibility priors, is performed to output the final logically coherent surgical action triplet.

To facilitate zero-shot surgical action triplet detection, we propose SurgZSD. As illustrated in Fig. 2, our framework localizes spatial regions via a fine-tuned Grounding DINO, performs medical knowledge-guided attribute composition to identify novel entities, and infers coherent triplets through joint inference constrained by clinical feasibility.

2.2 SFT-tuned Grounding DINO for Surgical Localization

To accurately obtain regions of interest (ROIs) for surgical instruments, we perform Supervised Fine-tuning (SFT) on Grounding DINO using the Cholec-Track20 [10] dataset. During the fine-tuning process, given an input surgical frame I and a specific text prompt T , the model is optimized by minimizing a combined localization and classification loss:

$$\mathcal{L}_{\text{SFT}} = \sum_i \left(\lambda_{\text{L1}} \mathcal{L}_{\text{L1}}(b_i, \hat{b}_i) + \lambda_{\text{giou}} \mathcal{L}_{\text{giou}}(b_i, \hat{b}_i) + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}(p_i, \hat{p}_i) \right) \quad (1)$$

where b_i and p_i denote the ground-truth bounding boxes and instrument category labels, while $\hat{b}_i$ and $\hat{p}_i$ represent the predicted box coordinates and class probabilities conditioned on the prompt T . The coefficients λ are hyper-parameters that balance the regression and classification tasks. At inference time, the fine-tuned Grounding DINO generates predicted instrument boxes and confidence scores, which serve as the spatial priors for the subsequent MKAC and CFCJI modules.

2.3 Medical Knowledge-Guided Attribute Composition Module

In Stage B, we convert the ROI localization obtained in Stage A into zero-shot category prediction, and perform interpretable inference over unseen entities by composing semantic attributes. To determine the discrete attributes of instruments, targets, and verbs within a given surgical frame I and its corresponding detection boxes $\{\hat{b}_i\}$, we leverage a pretrained VLM guided by structured prompts. These prompts are strictly formulated to compel the model to select exactly one valid option from a predefined candidate vocabulary for each attribute.

It is worth noting that the attribute dimensions are fixed for interpretability. Specifically, instruments, targets, and verbs are each described by three predefined attribute dimensions. For each branch, the candidate vocabulary of an attribute is constructed from the predefined attribute prototypes of all categories in that branch. Therefore, newly introduced categories can be incorporated by defining their corresponding attribute prototypes in the same attribute space.

Instrument attributes and scoring. We first construct a Knowledge Base (KB) generated by GPT-5.2, where each instrument category c is associated with a predefined attribute prototype $\mathbf{a}^c$. For each cropped RoI $x_i = \text{Crop}(I, \hat{b}_i)$, we query the specific instrument attributes $\{A_{\text{tip}}, A_{\text{joint}}, A_{\text{shaft}}\}$ to derive the predicted attribute vector $\hat{\mathbf{a}}_i^{(I)}$, and compute the final visual score for each instrument category via an exact matching criterion to output a Top- K candidate label set:

$$S_{\text{inst}}(i, c) = \sum_{k \in \{\text{tip}, \text{joint}, \text{shaft}\}} \mathbb{I}(\hat{a}_{i,k} = a_k^c), \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the condition holds and 0 otherwise. Thus $S_{\text{inst}}(i, c) \in \{0, 1, 2, 3\}$.

Target and verb attributes. Given the absence of precise spatial localization for targets and verbs under the zero-shot paradigm, we employ the instrument detection boxes as visual prompts to direct the VLM’s attention toward the areas of instrument-tissue contact. The framework independently queries the anatomical attributes $K_{\text{tar}} = \{\text{color}, \text{texture}, \text{morphology}\}$ alongside the verb attributes $K_{\text{ver}} = \{\text{tool_state}, \text{tissue_deformation}, \text{context_artifacts}\}$. By executing a weighted discrete matching against KB prototypes, we calculate the visual scores for target t and verb v over the i -th region:

$$S_{\text{tar}}(i, t) = \sum_{k \in K_{\text{tar}}} \mathbb{I}(\hat{a}_{i,k} = a_k^t), \quad S_{\text{ver}}(i, v) = \sum_{k \in K_{\text{ver}}} \mathbb{I}(\hat{a}_{i,k} = a_k^v) \quad (3)$$

Ultimately, rather than producing a single deterministic prediction, the framework outputs a Top- K candidate label set for both targets and verbs, accompanied by their corresponding visual scores. These candidate sets serve as the essential structured inputs for the subsequent joint inference module.

2.4 Clinical Feasibility-Constrained Joint Inference Module

In surgical environments, triplet elements are inherently interdependent and strictly governed by clinical feasibility. To ensure the prediction of clinically valid triplets, we propose a joint inference module that integrates attribute-composition scores with clinical feasibility priors.

Using the extracted Top- K candidate sets, we leverage GPT-5.2 to construct an Instrument-Target matrix $P_{IT}(c, t)$ and an Instrument-Verb matrix $P_{IV}(c, v)$ based on medical common sense. GPT-5.2 evaluates each pairwise relationship into three discrete feasibility levels:

$$P_{IT}(c, t), P_{IV}(c, v) \in \{\text{common}, \text{possible}, \text{impossible}\} \quad (4)$$

During joint inference, each feasibility level ℓ is mapped to an additive compatibility score $\phi(\ell)$. Specifically, we map the feasibility levels $\{\text{common}, \text{possible}, \text{impossible}\}$ to compatibility scores $\{+2, 0, -\gamma\}$ respectively, which are directly summed with the visual predictions. Accordingly, we define:

$$\phi_{IT}(c, t) = \phi(P_{IT}(c, t)), \quad \phi_{IV}(c, v) = \phi(P_{IV}(c, v)). \quad (5)$$

Let $S_{\text{inst}}(i, c)$, $S_{\text{tar}}(i, t)$, and $S_{\text{ver}}(i, v)$ denote the visual scores derived from the MKAC Module for the instrument, target, and verb, respectively. For each branch, we retain the top- K candidates and jointly infer the optimal triplet $(\hat{c}, \hat{t}, \hat{v})$ by maximizing the following objective:

$$\text{Score}(c, t, v) = S_{\text{inst}}(i, c) + S_{\text{tar}}(i, t) + S_{\text{ver}}(i, v) + \phi_{IT}(c, t) + \phi_{IV}(c, v) \quad (6)$$

This maximization is resolved via an exhaustive search over the retained top- K candidate sets.

3 Experiments and Results

3.1 Datasets and Experimental Details

We fine-tune Grounding DINO [7] on CholecTrack20 [10] for instrument localization. For evaluation, we utilize the official CholecT50 challenge [11] test set (1128 images) for closed-set scenarios. To assess zero-shot performance, we manually annotated 1377 images from ProstaTD [4] featuring instruments, targets, and verbs entirely absent from CholecT50, establishing a rigorous benchmark for novel entity recognition.

We report mean Average Precision (mAP), mean Average Recall (mAR), and Accuracy (Acc) as our primary evaluation metrics. Specifically, mAR measures the recall of instrument localization. For action triplet detection, following the CholecTriplet benchmark [9], a prediction is considered a true positive for mAP only if it achieves an Intersection over Union (IoU) > 0.5 with the ground-truth box and correctly identifies the entire triplet. Finally, Acc serves as a strict exact-match metric, measuring the proportion of predictions where the instrument, target, and verb are simultaneously recognized correctly. We adopt Qwen3VL-plus as the pretrained VLM. Our model was trained for 10 epochs on ten NVIDIA RTX 2080 Ti GPUs with a learning rate of 1×10^{-4} .

3.2 Results

Table 1 summarizes the quantitative zero-shot evaluation. Grounding DINO is unable to recognize action triplets, while SurgVLP [22] and Qwen3VL-plus struggle to output spatial localization. Therefore, only the instrument recognition accuracy is reported for SurgVLP and Qwen3VL-plus. In contrast, SurgZSD achieves superior instrument localization with 36.24% mAP and 73.90% mAR. Furthermore, it outperforms existing baselines in zero-shot surgical action triplet detection, reaching 6.42% mAP_{ivt} and significantly improving Acc to 11.33%. Qualitative results are illustrated in Figure 3.

Table 1: Zero-Shot Performance Comparison

Models	Instrument localization		Action triplet detection	
	mAP $\uparrow$	mAR $\uparrow$	Acc $\uparrow$	mAP_{ivt} $\uparrow$
Grounding DINO [7]	5.70%	35.60%	-	-
SurgVLP [22]	2.44%	-	0.04%	4.28%
Qwen3VL-plus	29.97%	-	5.68%	5.69%
SurgZSD(Ours)	36.24%	73.90%	11.33%	6.42%

As shown in Table 2, we further evaluate SurgZSD in closed-set scenarios. Notably, our framework achieves a superior instrument localization mAP of 65.39%. Although the triplet mAP_{ivt} of 4.68% trails specialized architectures, SurgZSD

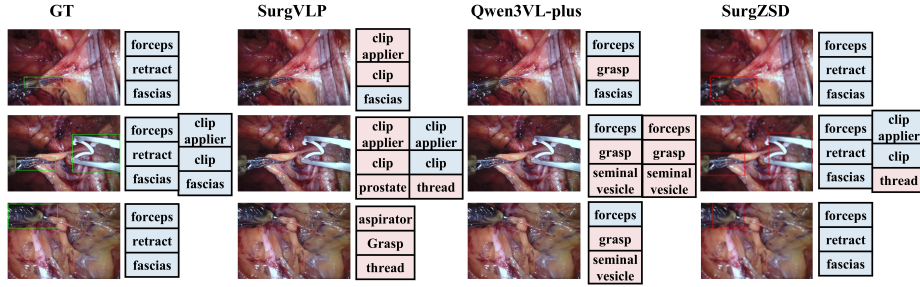


Fig. 3: **Qualitative results under zero-shot setting.** Green and red denote the correct and incorrect predictions, respectively.

maintains competitive closed-set performance without sacrificing its unique zero-shot generalization capabilities.

Table 2: Closed-Set Performance Comparison. Due to closed-source implementations, results for the two baselines are cited directly from their original papers.

Models	Instrument localization	Action triplet detection
	mAP $\uparrow$	mAP_{iwt} $\uparrow$
ITG-Trip [13]	59.10%	10.16%
MCIT-IG [18]	60.10%	7.32%
SurgZSD (Ours)	65.39%	4.68%

3.3 Ablation Studies

Table 3 details our step-wise ablation study. Removing the CFCJI module degrades performance, emphasizing the need for clinical constraints. Further omitting the MKAC module confirms its critical role in attribute parsing. Integrating both modules achieves the highest performance, proving the essential synergy between visual composition and clinical feasibility.

4 Conclusion

In this paper, we presented SurgZSD, a novel zero-shot surgical action triplet detection framework. By formulating scene understanding as a medical knowledge-guided and clinical feasibility-constrained joint inference process, our method effectively recognizes novel combinations of instruments, targets, and verbs. In future work, we plan to integrate temporal information and fine-tune a surgical-domain VLM to further enhance triplet recognition accuracy.

Table 3: Ablation study of the proposed components in SurgZSD

Settings	Instrument localization	Action triplet detection	
	mAP $\uparrow$	Acc $\uparrow$	$mAP_{ivt} \uparrow$
w/o MKAC & CFCJI	33.37%	6.84%	5.72%
w/o CFCJI	36.21%	8.93%	6.24%
SurgZSD (Ours)	36.24%	11.33%	6.42%

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 62322308, Grant 62133004, Grant 72188101, and Grant 72293585; in part by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM109; and in part by the Scientific Research Innovation Capability Support Project for Young Faculty under Grant SRICSPYF-ZY2025006.

Disclosure of Interests. The authors declare no competing interests.

References

1. Bai, L., Islam, M., Ren, H.: Cat-vil: Co-attention gated vision-language embedding for visual question localized-answering in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 397–407. Springer (2023)
2. Bai, L., Islam, M., Seenivasan, L., Ren, H.: Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. arXiv preprint arXiv:2305.11692 (2023)
3. Carstens, M., Vasisht, S., Zhang, Z., Barbur, I., Reinke, A., Maier-Hein, L., Hashimoto, D.A., Kolbinger, F.R.: Artificial intelligence for surgical scene understanding: a systematic review and reporting quality meta-analysis. npj Digital Medicine (2025)
4. Chen, Y., Li, Z., Xu, C., Liu, A.Q., Cui, R., Xu, X., Teoh, J.Y.C., He, S., Qin, J.: Prostatd: Bridging surgical triplet from classification to fully supervised detection. arXiv preprint arXiv:2506.01130 (2025)
5. Hao, P., Wang, H., Yang, G., Zhu, L.: Enhancing visual reasoning with llm-powered knowledge graphs for visual question localized-answering in robotic surgery. IEEE Journal of Biomedical and Health Informatics (2025)
6. Kitaguchi, D., Takeshita, N., Hasegawa, H., Ito, M.: Artificial intelligence-based computer vision in surgery: Recent advances and future perspectives. Annals of gastroenterological surgery **6**(1), 29–36 (2022)
7. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
8. Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Open world compositional zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5222–5230 (2021)

9. Nwoye, C.I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T., Jia, F., Yang, Y., Wang, H., et al.: Choelectriple2021: A benchmark challenge for surgical action triplet recognition. *Medical Image Analysis* **86**, 102803 (2023)
10. Nwoye, C.I., Elgohary, K., Srinivas, A., Zaid, F., Lavanchy, J.L., Padoy, N.: Choelectriple20: A multi-perspective tracking dataset for surgical tools. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8942–8952 (2025)
11. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis* **78**, 102433 (2022)
12. Nwoye, C.I., Yu, T., Sharma, S., Murali, A., Alapatt, D., Vardazaryan, A., Yuan, K., Hajek, J., Reiter, W., Yamlaoui, A., et al.: Choelectriple2022: Show me a tool and tell me the triplet—an endoscopic vision challenge for surgical action triplet detection. *Medical Image Analysis* **89**, 102888 (2023)
13. Pei, J., Zhang, J., Qin, G., Wang, K., Jin, Y., Heng, P.A.: Instrument-tissue-guided surgical action triplet detection via textual-temporal trail exploration. *IEEE transactions on medical imaging* (2025)
14. Poudel, N., Simon, R., Linte, C.A.: Evaluating large vision-language models for surgical tool detection. *arXiv preprint arXiv:2601.16895* (2026)
15. Qu, Y., Li, L., Li, T., Zheng, J., Tang, W., Li, X.: Smart surgical telementoring system with patch-to-pixel point tracking method. *EngMedicine* **2**(1), 100058 (2025)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
17. Sharma, S., Mutter, D., Padoy, N.: fine-clip: Enhancing zero-shot fine-grained surgical action recognition with vision-language models. *arXiv preprint arXiv:2503.19670* (2025)
18. Sharma, S., Nwoye, C.I., Mutter, D., Padoy, N.: Surgical action triplet detection by mixed supervised learning of instrument-tissue interactions. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 505–514. Springer (2023)
19. Su, Z., Guo, J., Yang, X., Wang, Q., Coenen, F., Huang, K.: Navigating distribution shifts in medical image analysis: A survey. *arXiv preprint arXiv:2411.05824* (2024)
20. Williams, S.C., Zhou, J., Muirhead, W.R., Khan, D.Z., Koh, C.H., Ahmed, R., Funnell, J.P., Hanrahan, J.G., Ali, A.M., Ghosh, S., et al.: Artificial intelligence assisted surgical scene recognition: a comparative study amongst healthcare professionals. *Annals of surgery* pp. 10–1097 (2024)
21. Williamson, T., Song, S.E.: Robotic surgery techniques to improve traditional laparoscopy. *JSL: Journal of the Society of Laparoscopic & Robotic Surgeons* **26**(2), e2022–00002 (2022)
22. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J.L., Marescaux, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures. *Medical Image Analysis* p. 103644 (2025)
23. Zhang, J., Zhou, S., Wang, Y., Zhao, H., Ding, H.: Knowledge-driven framework for anatomical landmark annotation in laparoscopic surgery. *IEEE transactions on medical imaging* **44**(5), 2218–2229 (2025)