Injecting Textual Spatial Context into Vision-Language Models for Surgical Scene Understanding

Antonio Martignano[1] ANTONIO.MARTIGNANO@SLU.EDU
Lin Guo[1] LIN.GUO@SLU.EDU
Chiara Camerota[1] CHIARA.CAMEROTA@SLU.EDU
[1] *Computer Science Department, Saint Louis University, USA*
Alessio Sacco[2] ALESSIO_SACCO@POLITO.IT
[2] *Department of Control and Computer Engineering, Politecnico di Torino, Italy*
Flavio Esposito[1] FLAVIO.ESPOSITO@SLU.EDU

Editors: Under Review for MIDL 2026

Abstract

Accurate anatomical landmark identification is important for safe laparoscopic navigation, yet limited view and strong tissue similarity make multi-label organ classification difficult. Existing vision-language models mainly rely on appearance and overlook the spatial structure of surgical scenes (Zhang et al., 2025). We propose *SpatialContext*, a multimodal framework that injects scene geometry into classification through natural language prompts derived from segmentation masks, together with a context-conditional training strategy centered on the primary surgical target. Results on DSAD (Carstens et al., 2023) and Endoscapes (Mascagni et al., 2025) show improved recognition of scene-defining and off-target anatomy, suggesting that explicit spatial semantics can improve surgical scene understanding.

Keywords: Laparoscopic Surgery, Vision-Language Models, Spatial Awareness.

1. Introduction

Laparoscopic surgeons operate from a restricted two-dimensional view, where visually similar tissues can make critical anatomy difficult to identify and increase the risk of iatrogenic injury (Kim et al., 2023). Although computer vision has strong potential as an intraoperative guide (Maier-Hein et al., 2022), dense supervision remains expensive in surgery (Willemink et al., 2020). As a result, datasets such as DSAD (Carstens et al., 2023) provide one dense primary-organ mask per image together with weak labels for additional visible structures, enabling scalable but challenging multi-label recognition. Standard detectors based on bounding boxes, such as YOLO (Redmon et al., 2016), are often too coarse for irregular anatomy, while weakly supervised classifiers still struggle with strong inter-tissue similarity.

Vision-language models such as CLIP (Radford et al., 2021) and BiomedCLIP (Zhang et al., 2025) offer an appealing alternative by aligning images and text without requiring closed-set classifiers, and BiomedCoOp (Koleilat et al., 2025) further improves this paradigm through learnable biomedical prompts. However, these methods primarily rely on visual appearance and do not explicitly model the spatial organization of surgical scenes. Graph-based approaches attempt to encode anatomy through scene relationships (Yuan et al., 2024), but they depend on reliable object detection and can fail when key structures are

not detected. We therefore propose *SpatialContext*, an initial step toward injecting explicit spatial cues into surgical vision-language modeling through natural language derived from the primary organ mask, with the broader goal of moving from texture matching toward scene-aware anatomical understanding.

2. System Design

Figure 1 illustrates SpatialContext. Frozen BiomedCLIP image and text encoders extract visual features and encode a spatial prompt generated from the centroid of the primary organ mask (e.g., *"The liver is in the top-right"*). Their embeddings are fused by an MLP and passed to independent sigmoid heads trained with weighted binary cross-entropy, enabling multi-label prediction. To avoid oracle leakage, context-conditional training masks the primary organ from both predictions and labels, forcing the model to use it only as a spatial anchor for detecting secondary anatomy.

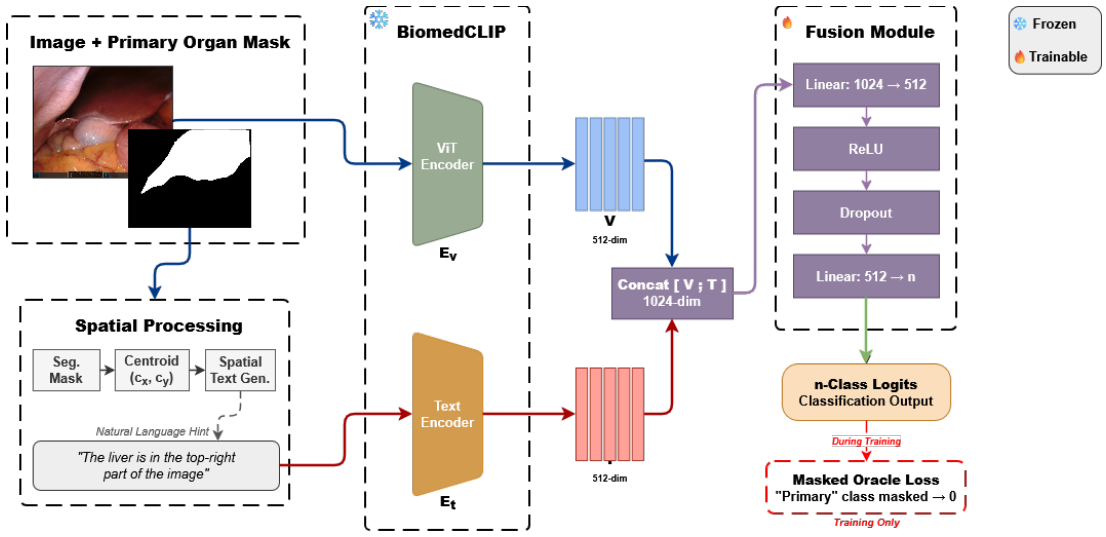

Figure 1: Overview of SpatialContext architecture.

3. Experiment and Results

We compare SpatialContext with zero-shot BiomedCLIP, BiomedCoOp, and a supervised YOLOv9 (Wang et al., 2025) baseline on DSAD and Endoscapes. To address class imbalance, we weight the positive term of binary cross-entropy by the class-specific ratio of negative to positive samples. We also sweep decision thresholds $\tau \in [0.001, 0.99]$ to distinguish detection errors from calibration errors.

Global performance. Table 1 shows that SpatialContext achieves 27.90% mAP on DSAD, outperforming BiomedCoOp, YOLOv9, and zero-shot BiomedCLIP. Figure 2 further shows that its gains are concentrated on large scene-defining organs such as the *Abdominal Wall* and *Small Intestine*, with additional improvement on the *Pancreas* and *Liver*, whereas BiomedCoOp remains stronger on finer localized structures including the *Inferior Mesenteric Artery (IMA)*, *Intestinal Veins*, and *Stomach*, suggesting that coarse spatial prompting mainly improves global scene understanding while prompt tuning better cap-

Table 1: Global performance under masked evaluation.

Dataset	Model	mAP	F1 ($\tau = 0.5$)	Opt. F1
DSAD	SpatialContext	**27.90%**	**45.23%**	**50.99%**
(11 classes)	BiomedCoOp	25.43%	15.23%	30.79%
	YOLOv9 (Expert)	13.15%	0.82%	23.47%
	BiomedCLIP (ZS)	12.64%	19.47%	19.47%
Endoscapes	SpatialContext	72.56%	27.54%	**80.89%**
(5 classes)	BiomedCoOp	**75.16%**	0.20%	80.45%
	YOLOv9 (Expert)	69.39%	10.47%	70.93%
	BiomedCLIP (ZS)	69.07%	**79.54%**	79.54%

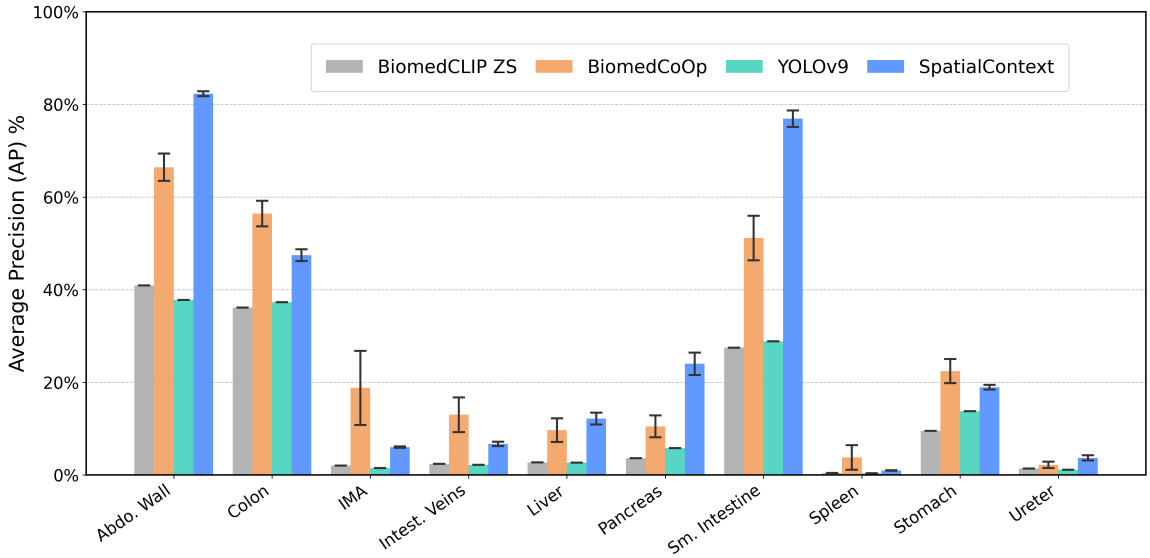

Figure 2: Per-class AP comparison on DSAD under masked evaluation.

tures fine-grained local appearance. **Scene vs. structure trade-off.** Spatial priors dominate on large, texture-poor regions that define the macroscopic scene: Abdominal Wall (AP 82.3%) and Small Intestine (AP 77.1%). For compact micro-anatomy defined by fine texture: ureter (AP 4.6%), spleen (AP 0.8%), BiomedCoOp retains superior performance. Spatial prompts introduce a low-resolution scene prior that regularises toward global layout at the cost of local detail. Prompt tuning conversely excels at high-resolution texture matching.

4. Conclusion

SpatialContext converts geometric centroids into natural language prompts, injecting coarse spatial context into a frozen VLM without dense graph annotation. Context-Conditional Training ensures genuine multimodal learning, while post-training threshold calibration remains important under severe class imbalance. Overall, this study provides an initial demonstration that coarse spatial prompting can improve scene-level anatomical understanding in weakly supervised surgical images. Future work will incorporate finer spatial detail and explicit inter-organ relations for more accurate spatial reasoning.

References

Matthias Carstens, Franziska M Rinner, Sebastian Bodenstedt, Alexander C Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R Kolbinger. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):3, 2023.

Jiyoun Kim, Yoon Jang, Su Hyeon Choi, Yong Wook Jung, Mi-La Kim, Bo Seong Yun, Seok Ju Seong, and Hye Sun Jun. Intraoperative fluorescent ureter visualization in complex laparoscopic or robotic-assisted gynecologic surgery. *Journal of Personalized Medicine*, 13(9), 2023. ISSN 2075-4426. doi: 10.3390/jpm13091345. URL https://www.mdpi.com/2075-4426/13/9/1345.

Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Biomedcoop: Learning to prompt for biomedical vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14766–14776, 2025.

Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. Surgical data science–from concepts toward clinical translation. *Medical image analysis*, 76:102306, 2022.

Pietro Mascagni, Deepak Alapatt, Aditya Murali, Armine Vardazaryan, Alain Garcia, Nariaki Okamoto, Guido Costamagna, Didier Mutter, Jacques Marescaux, Bernard Dallemagne, and Nicolas Padoy. Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. *Scientific Data*, 12(1), February 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-04642-4. URL http://dx.doi.org/10.1038/s41597-025-04642-4.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 1–21, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72751-1.

Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.

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

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2025. URL https://arxiv.org/abs/2303.00915.