FASL-Seg: Anatomy and Tool Segmentation of Surgical Scenes

Anonymous Author(s)
Affiliation
Address
email

Abstract

Achieving accurate surgical scene segmentation is crucial for the development of deep-learning-based surgical training for robotic minimally invasive surgeries. However, current state-of-the-art (SOTA) models struggle to balance capturing high-level contextual and low-level edge features to achieve holistic scene segmentation. We propose a Feature-Adaptive Spatial Localization model (FASL-Seg), designed to capture features at multiple levels of detail through two distinct processing streams, namely a Low-Level Feature Projection (LLFP) and a High-Level Feature Projection (HLFP) stream, effectively capturing anatomy and tools. We evaluated FASL-Seg on surgical benchmarks EndoVis18 and EndoVis17. The FASL-Seg model achieves a mean Intersection over Union (mIoU) of 72.71% (+5% over SOTA) on holistic scene segmentation in EndoVis18. It also achieves a mIoU of 85.61% and 72.78% in EndoVis18/EndoVis17 tool segmentation, respectively, outperforming SOTA with consistent performance on various classes for anatomy and instruments and demonstrating the effectiveness of distinct processing streams for holistic scene segmentation.

1 Introduction

While robot-assisted surgeries are becoming more popular [11], the complexity of the robotic system can challenge novice surgeons [12], requiring extensive training. Deep-learning algorithms can assist in surgical training by accurately identifying structures in the surgical scene, allowing for constructive feedback and objective skill assessments [1]. While various transformer-based models were proposed for surgical segmentation[19, 5, 21, 3], they often incorporate multiple backbones or decoders, increasing model complexity. Furthermore, little work considered segmentation of both anatomy and precise tools, with few reporting per-class metrics [10, 15]. The reason for this is that challenges arise when annotating the holistic surgical scene due to variance in tool and anatomical representations. Tool edges are often extracted in earlier encoder blocks while anatomical and contextual features are extracted in later stages of the segmentation model [14]. Achieving effective detail preservation and enhanced anatomical representation necessitates distinct processing of these multiscale features. We propose a Feature-Adaptive Spatial Localization (FASL-Seg) model. It composes of a transformer-based multiscale segmentation architecture that adapts feature processing using separate streams for low and high-level features. This approach maintains information integrity, ensuring fine-grained and precise localization of surgical tools and anatomy. To this end, our contributions are as follows:

- We propose a novel HLFP and LLFP-powered multiscale segmentation architecture that captures variational features with enhanced contextual understanding,
- Our proposed method is the combination of low-level features and high-level features that include edge information from initial layers of the network, which results in better segmentation performance,

- Our proposed architecture aggregates HLFP and LLFP in combination with a shallow decoder for both tools and anatomy segmentation,
- The code and trained models will be available upon acceptance

2 Method

Model Backbone This section details the proposed architecture for FASL-Seg, utilizing the SegFormer backbone [22]. This transformer model features hierarchical encoder blocks and attention mechanisms that enhance feature extraction for semantic scene understanding, particularly in medical applications [5]. SegFormer also eliminates positional encoding, making our model resilient to variations in input frame resolutions from different surgical platforms. Thus, we integrate SegFormer as our model backbone.

LLFP Stream To preserve the details of the first and second encoder feature maps, we propose a Low-Level Feature Projection (LLFP) stream to process local information. Given a feature map output of the i^{th} encoder block, F_i. Firstly, the feature map is passed through a Point-wise Convolution (PWConv) layer, followed by a Batch Normalization (BatchNorm) and Leaky ReLU (LReLU); hereafter, we shall refer to this combination as a ConvBlock. The PWConv layer allows spatial dimensions to be maintained while refining the feature representations. This is aided by using a small kernel size of 1, which prevents excessive smoothing of the fine details. The output of this block is then passed to a Multi-Head Self Attention (MHSA) block described in [20], with Attention formally represented in equation 1, where Q, K, and V are query, key and value vectors, W represents the output weight matrix, and d_k the key dimension scaling factor:

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V \tag{1}$$

The feature map is passed as the query, key, and value. With several heads, the details represented in the feature map are enhanced, as multiple representations can be learned concurrently. Furthermore, with higher resolution feature maps, local and global dependencies can be captured by the MHSA block, and irrelevant noise can be removed. The output of this block can thus be represented as:

$$\hat{F}_i \quad = \quad MHSA(ConvBlock(F_i))) \tag{2}$$

HLFP Stream Given a feature map, F_i, output from the i^{th} encoder block. Like the LLFP, the feature map is passed to a chain of Conv blocks, preserving the extracted contextual features, while enabling compression of channel-wise features into fewer channels. However, nlike in LLFP, where attention is needed to minimize noise and enhance the detailed features, the HLFP does not use MHSA, as the extracted features are already high-level and capture the global context with little noise. The introduction of attention would thus compromise essential features present in this stage. This block's output can thus be represented by equation 3, where $ConvBlock$ is a PWConv layer followed by BatchNorm and LReLU.

$$\hat{F}_i = ConvBlock_{1..N}(F_i) \tag{3}$$

Final Architecture of Model Figure 1 shows a simplified representation of the LLFP and HLFP streams, including necessary interpolation to match the feature map sizes. Equation 4 presents the fusion of the processed multiscale features from the four streams, where $F_i, i \in \{1, 2, 3, 4\}$ corresponds to the processed output of encoder block i. $\hat{F}_{EM}$ are passed through a shallow decoder of four PWConv blocks with BatchNorm and LReLU, to enable weighted feature selection and channel compression. Interpolation is used to increase feature map size before a final Laplacian convolution is applied to produce the segmentation output. Figure 2 presents the proposed FASL-Seg architecture.

$$\hat{F}_{EM} = Concat(\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4) \tag{4}$$

3 Datasets and Experimental Setup

To evaluate the efficacy of the proposed model architecture, the model was trained on two main datasets, namely EndoVis17 [6] and EndoVis18 [2] Challenge Datasets. For EndoVis18, the model

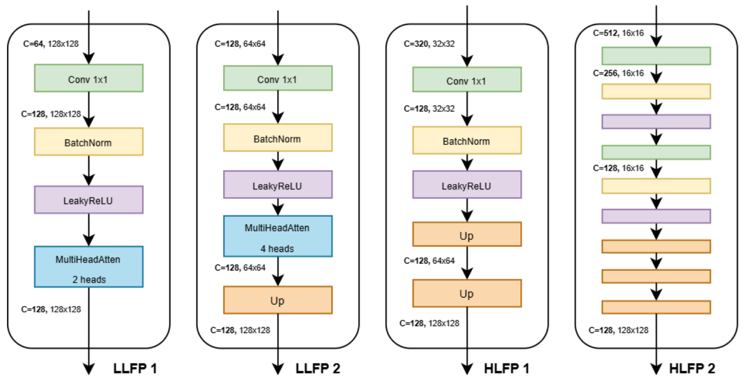

Figure 1: LLFP and HLFP streams in proposed architecture

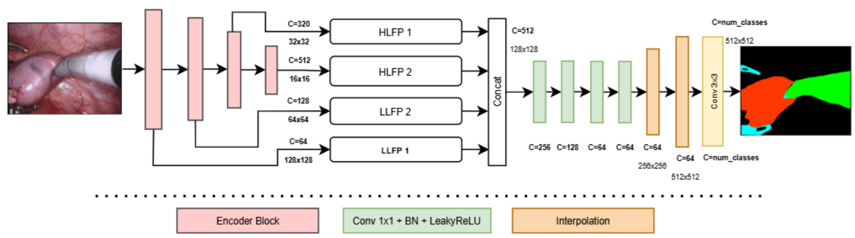

Figure 2: Overview of Proposed Architecture for FASL-Seg

was tested on holistic scene segmentation and tool segmentation, while only tool segmentation usecase exists in EndoVis17. Image sizes were 1280×1024, and official challenge train/test splits were used for all usecases (For EndoVis18 Tools, we followed the approach of González et al. [8]). Training sets were split into train/validation sets using a fixed seed for all experiments. All models and baselines were trained in a Linux environment utilizing an A6000 GPU with 48 GB GPU RAM, 8 CPUs and 45 GB of RAM. Models were trained for 100 epochs (approx. 24 hours), with a batch size of 4 due to memory constraints, and using an Adam optimizer with a fixed learning rate of 1E-5 for convergence stability, with no weight-decay. For the loss function, a combination of Tversky ($L_{tversky}$) [17] and Cross Entropy loss (L_{CE}) [13] was used, following equation 5, where α and β are set to 0.5. Performance was measured using mean Intersection over Union (mIoU) [8] and Dice Similarity Coefficient. All predictions were resized to original image sizes for measuring performance to ensure precision in quantifying performance on small or thin objects, often obscured upon resizing.

$$L_{total} = \alpha L_{tversky} + (1 - \alpha) L_{CE} \tag{5}$$

4 Results

Tables 1 and 2 show the performance of FASL-Seg on EndoVis18 Holistic segmentation and Tool Segmentation, respectively. In Holistic Segmentation, FASL-Seg raises mIoU from 0.65 to 0.73 (+8%) compared to SOTA. It also achieves highest average per-class mIoU, indicating more consistent segmentation across the different classes, thanks to the HLFP and LLFP streams. Similar performance is observed in Tool Segmentation, where FASL-Seg raises mIoU by 1% and Dice by 4% compared to SOTA on EndoVis18. It achieves top average per-class metrics, further showcasing consistency on various object representations. Additional results on EndoVis18 and EndoVis17 have been added to Appendix A.1, and further analysis of performance gain of FASL-Seg over SegFormer are presented in Appendix A.2. Figure 3 shows visual inference of FASL-Seg compared to various models on a set of holistic scene segmentation tasks. A qualitative analysis of the figures reveals the variance in performance across anatomy and tools for the baseline models. For instance, SegFormer segments anatomy well, but struggles with tool tips, whereas FASL-Seg consistently excels in segmenting both anatomical structures and tools. Its dedicated processing streams and shallow decoder enable it to retain essential details, making it effective for holistic scene segmentation tasks.

Table 1: Mean IoU Results EndoVis18 Parts and Anatomy Segmentation. Label Key: BT: Background Tissue ISh:Instrument Shaft, IC: Instrument Clasper, IW: Instrument Wrist, KP:Kidney Parenchyma, CK: Covered Kidney, SmInst: Small Intestine, SI: Suction Instrument, UP: Ultrasound Probe

Model	**mIoU**	**BT**	**ISh**	**IC**	**IW**	**KP**	**CK**	**Thread**	**Clamps**	**Needle**	**SI**	**SmInt**	**UP**	**Avg**
U-Net	0.53	0.65	0.72	0.37	0.45	0.43	0.07	0.38	0.76	0.91	0.70	0.18	0.72	0.53
Mask-RCNN	0.37	0.68	0.82	0.40	0.56	0.64	0.18	0.03	0.45	0.00	0.00	0.43	0.22	0.37
DeepLabV3	0.35	**0.87**	0.49	**0.62**	**0.72**	0.26	0.19	0.41	0.00	0.00	0.36	0.35	0.00	0.36
SegFormer	0.57	0.48	0.15	0.12	0.18	0.05	**0.52**	**0.90**	0.93	0.91	**1.00**	0.77	**0.85**	0.57
TransUNet [15]	0.48	0.59		0.60		0.32	0.33	0.01	**1.00**	0.04	0.64	0.63	0.61	0.48
MedT [15]	0.65	0.40		0.54		0.16	0.44	0.82	0.96	0.90	0.61	**0.78**	0.84	0.65
FASL-Seg(Ours)	**0.73**	0.79	**0.85**	0.53	0.65	**0.66**	0.29	0.69	0.90	**0.91**	0.85	0.77	0.84	**0.73**

Table 2: Results for EndoVis18 Tool Segmentation. Per-Class metrics is presented in the form mIoU[Dice]. Label Key: BF: Bipolar Forceps, PF: Prograsp Forceps, LND: Large Needle Driver, SI: Suction Instrument, VS: Vessel Sealer, GR: Grasping Retractor, CA: Clip Applier, MCS: Monopolar Curved Scissors, UP: Ultrasound Probe

Model	**mIoU**	**Dice**	**BF**	**PF**	**LND**	**SI**	**CA**	**MCS**	**UP**	**Avg**
UNet	0.64	0.66	0.66[0.74]	0.18[0.19]	0.48[0.50]	0.74[0.75]	0.8[0.8]	0.67[0.72]	0.62[0.62]	0.59[0.62]
SegFormer	0.71	0.72	0.11[0.11]	**0.87[0.87]**	0.82[0.82]	0.85[0.85]	**0.95[0.95]**	0.27[0.27]	**0.97[0.97]**	0.69[0.69]
TraSeTR [23]	-	-	0.76	0.53	0.47	0.41	0.14	0.86	0.18	0.48
S3Net [4]	0.74	-	0.77	0.51	0.20	0.51	0.0	0.92	0.07	0.48
MSLRGR [18]	-	-	0.70	0.44	0.0	0.35	0.04	0.87	0.12	0.36
MATIS [3]	0.84	-	0.82	0.47	0.66	0.69	0.00	0.92	0.21	0.54
ViTxCNN [21]	0.85	0.83	**0.86**	0.68	0.73	**0.89**	0.06	0.91	0.22	0.64
FASL-Seg(Ours)	**0.86**	**0.87**	0.79[**0.85**]	0.70[0.71]	**0.84[0.84]**	0.80[0.81]	0.95[0.95]	**0.92[0.95]**	0.88[0.88]	**0.84[0.85]**

5 Conclusion

FASL-Seg shows high potential in endoscopic datasets, yet further analysis is still required to assess the model's performance on cross-procedure generalisability. The use of Multi-Head Self Attention may also introduce computational complexities. Considering FASL-Seg's broader impact, although it aims to improve surgical situational awareness, its real-time use poses socio-technical risks, including over-trust by surgeons leading to oversight of errors, and false-predictions resulting in unnecessary cautery and instrument collisions, introducing ethical concerns. To address these issues, further clinical testing will include confidence maps, human-in-the-loop system to toggle overlays, and compliance with hazard analysis and surgical training. Cross-procedure datasets will be explored, in addition to lightweight attention mechanisms. Overall, FASL-Seg shows high potential for clinical integration in surgical training systems, and presents an important contribution to the field of holistic scene segmentation.

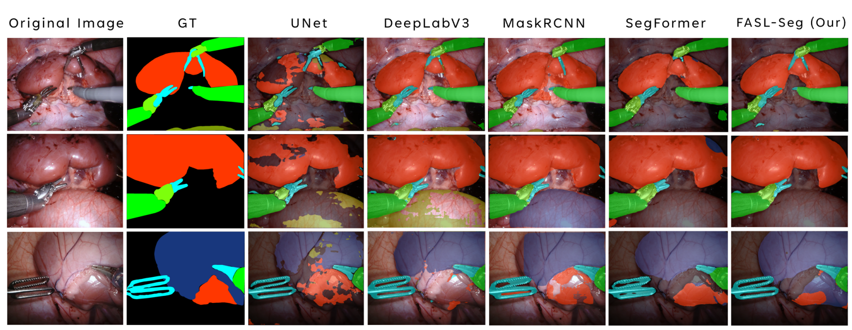

Figure 3: Inference Comparison on EndoVis18 Parts and Anatomy Segmentation of FASL-Seg against SOTA

References

[1] Ahmed, F. A., Yousef, M., Ahmed, M. A., Ali, H. O., Mahboob, A., Ali, H., Shah, Z., Aboumarzouk, O., Al Ansari, A., Balakrishnan, S., *Deep Learning for Surgical Instrument Recognition and Segmentation in Robotic-Assisted Surgeries: A Systematic Review*, Artificial Intelligence Review, Vol. 58, No. 1, 2025. DOI: 10.1007/s10462-024-10979-w. Available at: https://link.springer.com/article/10.1007/s10462-024-10979-w.

[2] Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., et al., *2018 Robotic Scene Segmentation Challenge*, In *International Journal of Computer Assisted Radiology and Surgery*, Springer, 2020. DOI: https://doi.org/10.1007/s11548-020-02196-9. Available at: https://link.springer.com/article/10.1007/s11548-020-02196-9.

[3] Ayobi, N., Pérez-Rondón, A., Rodríguez, S., Arbeláez, P., *MATIS: Masked-Attention Transformers for Surgical Instrument Segmentation*, Proceedings of the 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp. 1–5, 2023. DOI: 10.1109/ISBI53787.2023.10230819. Available at: https://ieeexplore.ieee.org/document/10230819.

[4] Baby, B., Thapar, D., Chasmai, M., Banerjee, T., Dargan, K., Suri, A., Banerjee, S., Arora, C., *From Forks to Forceps: A New Framework for Instance Segmentation of Surgical Instruments*, In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023. DOI: https://doi.org/10.48550/arXiv.2211.16200. Available at: https://openaccess.thecvf.com/content/WACV2023/papers/Baby_From_Forks_to_Forceps_A_New_Framework_for_Instance_Segmentation_WACV_2023_paper.pdf.

[5] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., Zhou, Y., *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*, In *Medical Image Analysis*, vol. 78, Elsevier, 2021. DOI: https://doi.org/10.1016/j.media.2021.102374. Available at: https://www.sciencedirect.com/science/article/pii/S1361841521001134.

[6] Duggal, R., et al., *Robotic Instrument Segmentation Challenge*, In *Proceedings of the MICCAI Endoscopic Vision Challenge*, MICCAI, 2017. Available at: https://github.com/duggalrahul/MICCAI17_EndoVis_RoboSeg.

[7] Gildenblat, J., and contributors, *PyTorch library for CAM methods*, In *GitHub Repository*, 2021. Available at: https://github.com/jacobgil/pytorch-grad-cam.

[8] González, C., Bravo-Sánchez, L., Arbelaez, P., *ISINet: An Instance-Based Approach for Surgical Instrument Segmentation*, In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2020. DOI: https://doi.org/10.1007/978-3-030-59716-0_57. Available at: https://link.springer.com/chapter/10.1007/978-3-030-59716-0_57.

[9] He, K., Gkioxari, G., Dollár, P., Girshick, R., *Mask R-CNN*, In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969. DOI: https://doi.org/10.1109/ICCV.2017.322. Available at: https://openaccess.thecvf.com/content_ICCV_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html.

[10] Jin, Y., Yu, Y., Chen, C., Zhao, Z., Heng, P.-A., Stoyanov, D., *Exploring Intra- and Inter-Video Relation for Surgical Semantic Scene Segmentation*, IEEE Transactions on Medical Imaging, Vol. 41, No. 11, pp. 2991–3002, 2022. DOI: 10.1109/TMI.2022.3177077. Available at: https://ieeexplore.ieee.org/document/9779714.

[11] Liu, M., Han, Y., Wang, J., Wang, C., Wang, Y., Meijering, E., *LSKANet: Long Strip Kernel Attention Network for Robotic Surgical Scene Segmentation*, IEEE Transactions on Medical Imaging, Vol. 43, No. 4, pp. 1308–1322, 2024. DOI: 10.1109/TMI.2023.3335406. Available at: https://ieeexplore.ieee.org/document/10302288.

[12] Matasyoh, W., Muthoni, J., and Zhang, W., *Samsurg: Surgical Instrument Segmentation in Robotic Surgeries Using Vision Foundation Model*, In *IEEE Access*, IEEE, 2024, Volume 12, Pages 12345–12356. DOI: https://doi.org/10.1109/ACCESS.2024.1234567. Available at: https://ieeexplore.ieee.org/document/1234567.

[13] Mao, A., Mohri, M., Zhong, Y., *Cross-Entropy Loss Functions: Theoretical Analysis and Applications*, In *Proceedings of the International Conference on Machine Learning (ICML)*, PMLR, 2023. Available at: https://dblp.org/rec/conf/icml/MaoM023a.

[14] Pang, Y., Li, Y., Shen, J., Shao, L., *Towards Bridging Semantic Gap to Improve Semantic Segmentation*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019. Available at: `https://openaccess.thecvf.com/content_ICCV_2019/papers/Pang_Towards_Bridging_Semantic_Gap_to_Improve_Semantic_Segmentation_ICCV_2019_paper.pdf`.

[15] Paranjape, J.N., Nair, N.G., Sikder, S., Vedula, S.S., Patel, V.M., *AdaptiveSAM: Towards Efficient Tuning of SAM for Surgical Scene Segmentation*, Medical Image Understanding and Analysis (MIUA 2024), Lecture Notes in Computer Science, Vol. 14860, Springer, Cham, pp. 187–201, 2024. DOI: `10.1007/978-3-031-66958-3_14`. Available at: `https://link.springer.com/chapter/10.1007/978-3-031-66958-3_14`.

[16] Ronneberger, O., Fischer, P., Brox, T., *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science, Vol. 9351, pp. 234–241, 2015. DOI: `10.1007/978-3-319-24574-4_28`. Available at: `https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28`.

[17] Salehi, S. S. M., Erdogmus, D., Gholipour, A., *Tversky loss function for image segmentation using 3D fully convolutional deep networks*, In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, 2017. DOI: `https://doi.org/10.1007/978-3-319-67389-9_44`. Available at: `https://link.springer.com/content/pdf/10.1007/978-3-319-67389-9_44.pdf`.

[18] Seenivasan, L., Mitheran, S., Islam, M., Ren, H., *Global-Reasoned Multi-Task Learning Model for Surgical Scene Understanding*, IEEE Robotics and Automation Letters, Vol. 7, No. 2, pp. 3858–3865, 2022. DOI: `10.1109/LRA.2022.3146544`. Available at: `https://ieeexplore.ieee.org/document/9695281`.

[19] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., Patel, V. M., *Medical Transformer: Gated Axial-Attention for Medical Image Segmentation*, In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2021*, Lecture Notes in Computer Science, vol. 12901, Springer, pp. 36–46, 2021. DOI: `https://doi.org/10.1007/978-3-030-87193-2_4`. Available at: `https://link.springer.com/chapter/10.1007/978-3-030-87193-2_4`.

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., *Attention Is All You Need*, Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, 2017. DOI: `10.5555/3295222.3295349`. Available at: `https://dl.acm.org/doi/10.5555/3295222.3295349`.

[21] Wei, M., Shi, M., Vercauteren, T., *Enhancing Surgical Instrument Segmentation: Integrating Vision Transformer Insights with Adapter*, International Journal of Computer Assisted Radiology and Surgery, vol. 19, pp. 1313–1320, 2024. DOI: `https://doi.org/10.1007/s11548-024-03140-z`. Available at: `https://link.springer.com/article/10.1007/s11548-024-03140-z`.

[22] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers*, Advances in Neural Information Processing Systems (NeurIPS), Vol. 34, pp. 12077–12090, 2021. DOI: `10.5555/3540261.3541185`. Available at: `https://dl.acm.org/doi/10.5555/3540261.3541185`.

[23] Zhao, Z., Jin, Y., Heng, P.-A., *TraSeTR: Track-to-Segment Transformer with Contrastive Query for Instance-level Instrument Segmentation in Robotic Surgery*, In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2022. DOI: `10.1109/ICRA46639.2022.9811873`.

A Technical Appendices and Supplementary Material

Technical appendices with additional results, figures, graphs and proofs may be submitted with the paper submission before the full submission deadline (see above), or as a separate PDF in the ZIP file below before the supplementary material deadline. There is no page limit for the technical appendices.

A.1 Additional Results on Segmentation Datasets

Table 3 presents the performance in Dice Similarity Coefficient of FASL-Seg against SOTA on EndoVis18 Holistic Segmentation. Similar to the performance in mIoU, FASL-Seg achieves new SOTA performance and higher consistency across various classes of anatomy and tools. Similarly, FASL-Seg achieves top results on EndoVis17 Tool Segmentation, as shown in Table 4.

Table 3: Dice Results for EndoVis18 Parts and Anatomy Segmentation. Label Key: BT: Background Tissue ISh:Instrument Shaft, IC: Instrument Clasper, IW: Instrument Wrist, KP:Kidney Parenchyma, CK: Covered Kidney, SmInst: Small Intestine, SI: Suction Instrument, UP: Ultrasound Probe

Model	Dice	BT	ISh	IC	IW	KP	CK	Thread	Clamps	Needle	SI	SmInt	UP	Avg
U-Net	0.58	0.78	0.78	0.50	0.55	0.54	0.11	0.38	0.77	0.91	0.70	0.20	0.73	0.58
Mask-RCNN	0.48	0.81	**0.90**	0.57	0.72	**0.78**	0.31	0.07	0.62	0.0	0.0	0.60	0.36	0.48
DeepLabV3	0.46	**0.93**	0.66	**0.76**	**0.84**	0.42	0.31	0.58	0.00	0.00	0.53	0.52	0.00	0.46
SegFormer	0.58	0.64	0.12	0.12	0.18	0.05	**0.52**	**0.90**	0.93	0.91	**1.00**	0.77	**0.85**	0.58
TransUNet [15]	0.52	0.73		0.70		0.49	0.33	0.00	**1.00**	0.04	0.66	0.63	0.62	0.52
MedT [15]	0.68	0.56		0.66		0.26	0.44	0.82	0.96	0.90	0.62	0.78	0.84	0.68
FASL-Seg(Ours)	**0.77**	0.87	0.89	0.65	0.74	0.72	0.34	0.71	0.91	**0.91**	0.85	**0.79**	**0.85**	**0.77**

Table 4: Results for EndoVis17 Tool Segmentation. Per-Class metrics is presented in the form mIoU[Dice]. Label Key: BF: Bipolar Forceps, PF: Prograsp Forceps, LND: Large Needle Driver, VS: Vessel Sealer, GR: Grasping Retractor, MCS: Monopolar Curved Scissors, UP: Ultrasound Probe

Model	mIoU	Dice	BF	PF	LND	VS	GR	MCS	UP	Avg
UNet	0.42	0.44	0.17[0.19]	0.16[0.18]	0.26[0.28]	0.39[0.39]	0.52[0.52]	0.63[0.66]	0.33[0.33]	0.35[0.36]
SegFormer	0.67	0.68	0.46[0.48]	0.44[0.46]	0.54[0.56]	0.63[0.63]	0.79[0.79]	0.68[0.7]	0.87[0.97]	0.63[0.66]
TraSeTR [23]	0.65	-	0.45	0.57	0.56	0.39	0.11	0.31	0.18	0.37
S3Net [4]	0.72	-	**0.75**	0.54	0.62	0.36	0.27	0.43	0.28	0.47
MATIS [3]	0.71	-	0.69	0.52	0.52	0.32	0.19	0.24	0.25	0.65
ViTxCNN [21]	0.69	-	0.66	**0.68**	**0.71**	0.43	0.13	0.40	0.29	0.47
FASL-Seg(Ours)	**0.73**	**0.74**	0.62[0.64]	0.54[0.55]	0.63[0.64]	**0.66[0.66]**	**0.81[0.81]**	**0.74[0.76]**	**0.89[0.89]**	**0.70[0.71]**

A.2 Analysis of FASL-Seg and SegFormer Performances

We further assess our model's predictions using False Positive Rate (FPR), which is defined as $\frac{FalsePositive+\epsilon}{FalsePositive+TrueNegative+\epsilon}$, where ϵ is set to 1E-7. The results for each dataset against the SegFormer-b5 model are shown in Table 5. The results show that FASL-Seg has improved FPR results over SegFormer at 1.1% in combined Parts and Anatomy classes, 1.04% for anatomy only and 1% in Tools segmentation. This is attributed to the improved holistic mask generation due to adaptive processing of thin tools as well as larger tools and anatomy through the distinct processing streams.

Table 5: False Positive Rate (FPR) results for the EndoVis18 Parts and Anatomy and Tool Segmentation datasets for FASL-Seg against SegFormer

Classes on EndoVis18	SegFormer-b5 (FPR)	FASL-Seg (FPR)
Parts and Anatomy Segmentation	0.078	**0.067**
Only Parts	0.003	0.003
Only Anatomy	0.0704	**0.06**
Tools Segmentation	0.005	**0.004**

A.3 Ablation Studies

To analyze the effectiveness of the proposed architecture, a thorough ablation study was conducted on the streams' components. First, we analyze the use of attention in the LLFPs and HLFPs, to assess the added value from the multi-head self attention on the feature representations extracted from each encoder output. The investigated configurations and corresponding model performances are presented in Table 6.

The results reveal that applying attention on all the hidden state projections does not necessarily improve the model segmentation ability. Applying attention on HLFP2, which projects the smallest feature map size, shows a drop in the performance compared to no attention applied. This is supported by the knowledge that later encoder blocks encode high-level features; applying attention may result in loss of crucial semantic understanding of the surgical scene. Contrastively, earlier encoder feature maps have more fine-grained knowledge of the image content. Thus, it is observed that attention applied to LLFP1 and LLFP2 improved the mean IoU and Dice

Table 6: Results for Ablation on Attention Utilization in Feature Processing Streams

Model	LLFP1	LLFP2	HLFP1	HLFP2	**mIoU**	**Dice**
Model-1					0.6679	0.7103
Model-2	✓				0.6785	0.7202
Model-3		✓			0.6716	0.7135
Model-4				✓	0.6669	0.7085
Model-5			✓	✓	0.6497	0.6909
FASL-Seg	✓	✓			**0.6823**	**0.7236**

Table 7: Ablation on Number of Attention Heads. For per-class, mIoU is presented. Label Key: ISh:Instrument Shaft, IC: Instrument Clasper, IW: Instrument Wrist, SI: Suction Instrument, UP: Ultrasound Probe

Model	Atten. Heads			**mIoU**	**Dice**	**ISh**	**IC**	**IW**	**Thread**	**Clamps**	**Needle**	**SI**	**UP**
	1	2	4										
Model-6	✓			0.6392	0.6803	0.82	0.522	0.585	0.22	0.74	0.905	0.754	0.772
Model-7		✓		0.6535	0.6938	0.832	0.526	0.595	0.311	0.78	0.905	0.746	0.751
Model-8			✓	0.6441	0.6845	0.79	0.528	0.597	0.288	0.798	0.905	0.735	0.772
FASL-Seg		✓	✓	**0.6823**	**0.7236**	**0.847**	**0.526**	**0.649**	**0.483**	**0.799**	**0.905**	**0.811**	**0.802**

compared to no attention. The next ablation study was conducted on the number of attention heads to use in the LLFPs. One, two, and four heads were investigated. The results are presented in Table 7. Initially, the overall performances reveals two head attention performs better than the one or four heads. However, per-class results reveal that some classes were captured better with four-head attention than with two-head and vice versa. An additional experiment was conducted with a combination of two-head attention in LLFP1 and four-head attention in LLFP2, which resulted in the best performance.

An ablation study was conducted on the use of Convolution Transpose (ConvTrans) layers against regular Bilinear interpolation. The ConvTrans were followed by batch normalization and ReLU matching the Conv Blocks used in the rest of the architecture. Surprisingly, using ConvTrans blocks did not provide additional insights on the feature maps, and lead to losing important insights when the features were enlarged. Instead, the use of interpolation was found to improve the model output.

A.4 FASL-Seg Model Complexity against SOTA

A comparison between the model complexity of FASL-Seg against several SOTA methods is presented in Table 9. Despite some SOTA architectures having more parameters or FLOPS, FASL-Seg was able to outperform them in the three benchmarks. Furthermore, FASL-Seg has lower parameters than SegFormer even with the additional components. The inference speed of our model on the A6000 GPU with 48GB of RAM was 2.14 frames per second, with peak GPU memory at 0.92GB. Thus, the current state of FASL-Seg is more suitable to run post-operative analysis of surgical videos.

Table 8: Results of Ablation on Upsampling Mechanism used throughout the model architecture

Model	ConvTrans	Interpolation	mIoU	Dice
Model-9	✓		0.6823	0.7236
FASL-Seg		✓	**0.7222**	**0.7622**

Table 9: Model Complexity of FASL-Seg against SOTA

Model	Architecture	#Params	GFLOPs
UNet	CNN	13.39M	124.44
MaskRCNN	CNN	44M	447
DeepLabV3	CNN	60.99M	258.74
TransUNet	Transformer,CNN	105.3M	38.6
SegFormer-b5	Transformer	84.6M	110.25
FASL-Seg	Transformer,CNN	81.99M	223.42

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In our work, we present a new transformer-based architecture to improve holistic scene segmentation of anatomical and instrument structures in a surgical scene. To validate this claim, we conduct experiments on holistic scene segmentation using the EndoVis18 dataset, as well as tool segmentation using EndoVis18 and EndoVis17 datasets. FASL-Seg raises performance on these datasets over SOTA and presents more consistent per-class metrics for anatomy and tools, supporting the claim that it improves holistic surgical scene segmentation.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In our conclusion, we discuss two limitations of the current work. Firstly, that only endoscopic surgeries were investigated, thus future work will explore cross-procedure datasets. Secondly, the use of Multi-Head Self Attention introduces computational overhead. Exploring more lightweight options can help reduce this overhead and introduce the model for real-time deployment. Detailed computational complexity discussions are presented in the Appendix due to limited number of pages for main manuscript.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theories were presented.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: To support reproduciblity, a detailed description of the proposed architecture is presented. The architecture diagram, Figure 2, presents detailed channel input and output values as well as feature map sizes, enabling any researcher to recreate the model from scratch. Experimental setup and datasets are also presented in detail in a dedicated section. Upon acceptance of the paper, a link to the code and model weights will be shared.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but

reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All data used in this work is publicly available online. Code is publishable upon acceptance of the paper, and the link will be shared in the paper. This is also to support anonymity of the submitted work.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details were provided in a dedicated section including learning rate, optimizer, loss functions, train/test splits, GPU hardware used, and other details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The overall and per-class metrics are reported following established surgical scene segmentation performance metrics. We focus on established metrics for comparison with the literature, namely mean Intersection over Union, Dice Similarity Coefficient, and False Positive Rate.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide details of the training hardware, number of epochs and batch sizes with approximate time consumption in the Experimental Setup and Datasets section. We also discuss runtime and peak memory consumption in the Appendix Model complexity section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our work adheres to all the guidelines in the Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: In our conclusion, we discuss the potential societal impacts and possible negative impacts if the model was used in real-time during a surgical procedure, and briefly discuss ways to address these issues during clinical analysis of the model.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No such model was developed.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All authors that have contributed to the code used for the experiments reported in this paper have been added as authors and coauthors to credit their work. All data that is publicly available has been properly cited to give credit to the original authors of those datasets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All code is well documented and prepared for publication upon acceptance of this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No such research was conducted.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No such research was conducted.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: As the LLM was only used for writing and editing, no declaration is required.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.