Projection-Based Quality Assessment of Cerebrovascular Segmentations

Bogdan Ion[1], Xiaoming Zhang[1[0009−0000−3986−885X]], Leonard Vincent Ramil[1,2], Alice Boccadifuoco[1,2], Sébastien Ourselin[3[0000−0002−5694−5340]] Jon Cleary[3[0000−0001−8836−9934]], and Maria A. Zuluaga[1,3[0000−0002−1147−766X]]

[1] EURECOM, Sophia Antipolis, France
[2] Politecnico di Torino, Turin, Italy
[3] School of Biomedical Engineering & Imaging Sciences, King's College London, UK
maria.zuluaga@eurecom.fr

Abstract. Reliable cerebrovascular analysis from 3D time-of-flight magnetic resonance angiography depends on the anatomical completeness and consistency of vascular representations. However, automated vessel segmentation may contain errors that hinder downstream analysis tasks. Identifying such errors directly in 3D is challenging due to the complex topology of the cerebrovascular anatomy. In this work, we propose a canonical multi-view framework for automated assessment of cerebrovascular segmentation quality. Rather than reasoning directly over the complete 3D vascular tree, we reformulate localized quality assessment as detection in canonical two-dimensional projection views, where anatomical structures exhibit reproducible appearance. A lightweight detection model is applied independently to complementary views, whose agreement is used to identify potentially erroneous segmentations for manual inspection. To alleviate annotation scarcity, we propose view synthesis by perturbing projection angles and slabs around anatomy-guided reference views, generating anatomically valid training examples from the same registered 3D volume. We demonstrate the framework for identifying spurious superior sagittal sinus segmentations in TOF-MRA and evaluate it on 99 subjects with manually annotated quality labels. The proposed approach achieved an F1 score of 88.31, outperforming vision-language models (VLM) operating in few-shot settings or adapted with LoRA, while relying on a lightweight task-specific detector instead of large VLMs. Our results show that canonical projection views reformulate 3D cerebrovascular quality assessment into a robust and efficient 2D detection problem, facilitating scalable construction of reliable cerebrovascular analysis pipelines. All code is available at github.com/erc-caravel/vascular-qc.

Keywords: Cerebrovasculature · Quality Assessment · Segmentation

1 Introduction

Three-dimensional time-of-flight magnetic resonance angiography (TOF-MRA) provides a non-invasive imaging modality for visualizing intracranial arteries

and is increasingly used for vessel segmentation, atlas construction, and downstream morphometric analysis [4]. However, their reliability depends critically on the anatomical fidelity of the underlying vascular representations. Missing vascular segments or spurious vessel segmentations can propagate into atlas estimation, bias population statistics, and reduce the clinical interpretability of derived biomarkers.

Quality assessment of cerebrovascular segmentations remains challenging. The cerebral vascular tree is thin, tortuous, and highly variable across subjects.Directly evaluating the full 3D vessel tree requires reasoning about complex topology and long-range anatomical relationships. Manual inspection by experts remains the most reliable approach, but it is time-consuming and difficult to scale to large cohorts. This creates a need for automated quality-control methods that can identify segmentation failures while remaining efficient and compatible with limited expert annotation.

A natural strategy is to exploit prior anatomical knowledge. Although the full cerebrovascular network is complex, individual structures may be assessed from canonical projections where they are consistently visualized. For example, the superior sagittal sinus (SSS) can be visualized from sagittal and oblique axial projections, where errors may be easier to recognize than in the original 3D volume. This observation motivates a reformulation of local vascular quality assessment from a global 3D reasoning problem into a set of 2D detection tasks. Such a formulation also enables the use of lightweight detection models trained on targeted annotations rather than requiring large-scale 3D labels.

In this work, we propose a canonical multi-view framework for automated cerebrovascular segmentation quality assessment. Subjects are first rigidly aligned to a common anatomical reference frame. Maximum intensity projections are then computed from predefined anatomical viewpoints selected to best visualize the vessels of interest. Lightweight detectors are applied independently to complementary views, whose agreement is used to flag potentially erroneous segmentations. To alleviate annotation scarcity, we further introduce view synthesis by perturbing the projection parameters around the predefined viewpoints.

The main contribution of this work is a framework for assessing the quality of cerebrovascular segmentation based on canonical projection views. Building on this representation, we combine complementary multi-view detection with projection-based view synthesis to obtain an efficient framework that requires only lightweight, task-specific detectors. We demonstrate the proposed approach for identifying spurious superior sagittal sinus (SSS) segmentations in TOF-MRA, highlighting its potential for scalable cerebrovascular quality assessment.

2 Related Works

Medical Image Segmentation Quality Assessment. Automated quality control (QC) has become an increasingly active research topic in medical image analysis, aiming to identify unreliable segmentations without requiring manual inspection. Recent work has explored learned representations to estimate seg-

mentation quality across different organs and imaging modalities [3,10]. Despite these advances, QC of vascular segmentations remains largely unexplored due to the complexity of vascular anatomy, characterized by thin, tortuous, and highly interconnected structures. Existing work has primarily focused on retinal vessel analysis from 2D fundus images [8], whereas quality assessment of three-dimensional cerebrovascular segmentations remains unexplored.

Vision-Language Models for Medical Image Understanding. Recent vision language models (VLMs), including InternVL3.5 [2] and Qwen2.5-VL [1], have demonstrated strong zero-shot and few-shot capabilities across a broad range of visual reasoning tasks. Their adaptation to medical imaging, through models such as LLaVA-Med [9], has extended these capabilities to clinical applications. Moreover, VLMs have demonstrated visual grounding capabilities, localizing objects or regions from natural language or visual prompts [1], a capability that has also been explored for locating anatomical findings from textual descriptions [11]. However, highly specialized vascular representations, such as vessel-only maximum intensity projections derived from TOF-MRA, remain substantially different from the visual distributions encountered during pre-training. As a result, effective deployment often requires task-specific adaptation, for example, through parameter-efficient fine-tuning methods such as LoRA [6].

Lightweight Object Detection for Anatomical Localization. Lightweight object detectors, such as the YOLO family [7], learn spatial appearance directly from annotated examples and achieve accurate localization/detection with substantially lower computational requirements than large foundation models. When anatomical variability is reduced through canonical representations, localization becomes a spatial recognition problem, making lightweight detectors an efficient alternative to general-purpose VLMs.

3 Method

Figure 1 illustrates the proposed framework. Starting from an MRA volume and its corresponding vessel segmentation, subjects are first aligned to a common anatomical reference frame to standardize anatomical orientation. Canonical maximum-intensity projections are then computed from predefined viewpoints selected to best visualize the vascular structures of interest. These projections are analyzed independently by a lightweight detector, and agreement across complementary views is used to identify potentially erroneous segmentations. To alleviate annotation scarcity, additional training views are synthesized by perturbing the projection parameters around the canonical viewpoints.

Projection Generation. Quality assessment of 3D vascular segmentations requires reasoning over a complex vascular network with substantial anatomical variability. Rather than assessing cerebrovascular segmentations directly in three dimensions, we reformulate quality assessment as the analysis of two-dimensional views of the segmented vasculature. The key idea is that projections computed from predefined viewpoints retain sufficient anatomical information to assess vascular structures while substantially reducing the complexity of the problem.

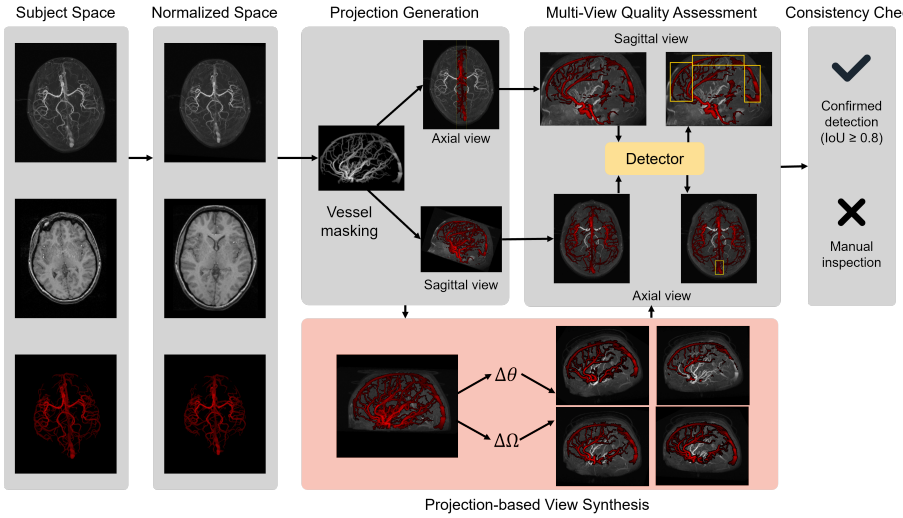

Fig. 1: Overview of the proposed multi-view quality assessment framework based on canonical projection views.

To ensure that the same projection parameters correspond to anatomically comparable viewpoints across subjects, all TOF-MRA volumes and their corresponding vessel segmentations are first rigidly registered to a common anatomical reference frame through the corresponding T1-weighted image. To isolate the vascular signal, the arterial segmentation is first applied as a binary mask over the corresponding MRA volume, preserving vessel intensities while suppressing surrounding anatomy $V_{\text{mask}} = V_{\text{MRA}} \odot S$, where $\odot$ denotes voxel-wise multiplication, V_{MRA} is the angiographic volume, and S is the corresponding arterial segmentation. The resulting masked volume preserves the original vessel intensities while suppressing non-vascular anatomy.

Maximum intensity projections (MIPs) are subsequently generated from the masked volume using a projection operator defined by the viewing orientation θ, and the projection slab Ω as $MIP = P(V_{\text{mask}}; \theta, \Omega)$, where the projection operator $P(\cdot)$ is defined as

$$P(V; \theta, \Omega) = \max_{w \in \Omega} V(R_\theta(u, v, w)), \tag{1}$$

where R_θ denotes a rotation of the volume by an angle θ, and Ω represents the selected projection slab.

Multi-View Quality Assessment. We consider the target segmentation protocol to include only arteries; thus, the presence of the SSS indicates a false positive error. Therefore, the quality assessment task consists of identifying the SSS from complementary projection views. To this end, generated projection views are analyzed independently using an object detector. In this work, the detector is trained to localize three predefined regions of the SSS, corresponding to

the anterior, central, and posterior regions, each represented as an independent detection class. Treating these regions as independent detection targets enables localized quality assessment while reducing the spatial search space.

The detector is applied to each complementary projection view generated from the registered volume. As the visibility of the SSS varies across projection directions, each detection is cross-verified using the complementary view. Specifically, a detection in one projection is mapped onto the complementary projection using the known projection geometry, thereby defining the region where the corresponding anatomical structure is expected to appear. The corresponding detection in the complementary view is then evaluated against this projected region. The expected location in the complementary view is deterministically derived from the same region of interest in the registered 3D volume, constraining the search space, thus a detection is confirmed when the predicted and projected regions overlap with a moderate Intersection-over-Union ≥ 0.5. Agreement between the two detections provides corroborating evidence of the presence of the SSS, whereas missing or geometrically inconsistent detections indicate an uncertain finding and trigger manual inspection. Verified detections are subsequently mapped back to the registered 3D volume, easing rapid visual verification of the segmentation errors.

View Synthesis for Data Augmentation. Manual annotations of cerebrovascular structures are scarce and costly to obtain. To alleviate this, we exploit the capacity to project an annotated volume into multiple projection views that are anatomically consistent with the original volume. As such, we increase the diversity of the training set without requiring additional 3D imaging data.

Rather than augmenting the resulting two-dimensional projections, additional views are synthesized directly from the registered 3D volume by perturbing the projection parameters around a predefined reference view. The perturbations are chosen to preserve target vessel visibility while introducing realistic variations in view orientation and projection thickness.

For a reference orientation θ and slab Ω, we define a set of augmented views as

$$MIP_{\text{aug}} = \{P(V^*_{\text{mask}}; \theta + \Delta\theta_k, \Omega_k)\}_{k=1}^{K}, \tag{2}$$

where $\Delta\theta_k$ denotes the angular perturbation, Ω_k the corresponding projection slab, and K the number of synthesized views. Because all subjects are represented in a common anatomical space, the same perturbation strategy can be applied consistently across the dataset.

4 Experimental Setup

Datasets and Experimental Split. We evaluated the proposed framework on the IXI and TubeTK datasets using manual segmentations from the VesselVerse dataset [5]. Two canonical projection views were defined for the SSS, a midsagittal view and an oblique axial view, selected to maximize visualization of the target vessel. The SSS annotations were manually created by one annotator and subsequently reviewed and validated by an expert. The training set

comprised 230 IXI subjects and 10 manually annotated TubeTK subjects, while the remaining 99 TubeTK subjects formed an independent test set. Each annotated TubeTK subject generated 15 synthesized projections, resulting in 150 training projections from the 10 annotated subjects. The test set contained 38, 35, and 5 positive instances for the anterior, posterior, and central SSS regions, respectively (78 in total).

We evaluated two training configurations to assess the impact of view synthesis. The baseline used a single projection from each of the 10 annotated TubeTK subjects (*IXI+10*), whereas the proposed approach generated 150 synthesized projections by perturbing the projection parameters (*IXI+150*). The IXI training set and the TubeTK test set remained unchanged.

Implementation details. We employ YOLOv11m as a lightweight object detector, initialized with COCO pre-trained weights and trained with bounding-box supervision. Models were trained independently for the *IXI+10* and *IXI+150* configurations using identical hyperparameters. Training was performed using AdamW optimizer for 300 epochs with a batch size of 8, an initial learning rate of 0.01, and an input size 448×448.

Baselines We compare our strategy to two types of VLMs. *1) Prompted VLMs.* InternVL3.5-14B, Qwen2.5-VL-72B, and Gemma 3-31B were evaluated using two-shot prompting, with demonstrations selected exclusively from the training set. Each model was prompted to identify the anterior, posterior, and central SSS regions. *2) LoRA-adapted VLM.* InternVL3.5-14B was additionally fine-tuned using LoRA on the *IXI+150* training configuration to evaluate whether task-specific adaptation could bridge the gap with lightweight detectors. "We restricted parameter-efficient fine-tuning to InternVL3.5-14B, since the available annotated data in this domain is limited, and adapting substantially larger models for the comparatively simple task of 2D pattern recognition would be computationally disproportionate to the expected gain.

Evaluation Protocol. All models were evaluated on the same 99 held-out TubeTK subjects. Predictions were assessed independently for each anatomical region using a one-versus-rest formulation. True positives (TP), false positives (FP), and false negatives (FN) were computed by matching predicted and ground-truth regions, with duplicate unmatched predictions counted as FPs.

For each SSS sub-region c, we estimate precision (P), recall (R), and F1 score. Macro-averaged metrics were obtained by averaging the corresponding class-specific values over the three SSS regions. Micro-averaged metrics were calculated after pooling TP, FP, and FN counts across classes. We additionally report the pooled detection Jaccard score, $J_{\text{det}} = \sum_c \text{TP}_c / (\sum_c \text{TP}_c + \sum_c \text{FP}_c + \sum_c \text{FN}_c)$.

5 Results

SSS Detection Benchmark. We compare our approach with prompted VLMs. Table 1 shows that prompted VLMs failed to achieve a balance between sensitivity and precision. Although the prompted VLMs exhibited markedly different

Table 1: SSS detection benchmark. Macro metrics average over the three SSS classes; micro metrics pool TP, FP, and FN across classes. All metrics except TP, FP, and FN are percentages. Best F1 and Jaccard scores are in bold.

Method	Setting	TP	FP	FN	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1	J_{det}	Size
Two-shot prompted VLM												
InternVL3.5-14B	2-shot	13	12	65	52.71	23.13	24.80	52.00	16.67	25.24	14.44	29 GB
Qwen2.5-VL-72B	2-shot	43	73	35	38.49	67.94	41.56	37.07	55.13	44.33	28.48	137 GB
Gemma 3-31B	2-shot	73	191	5	27.40	95.61	39.71	27.65	93.59	42.69	27.14	59 GB
Task-specific training												
YOLO	IXI+10	29	3	49	60.52	26.57	36.84	90.63	37.18	52.73	35.80	39 MB
InternVL3.5-14B LoRA	IXI+150	60	14	18	81.53	77.97	79.64	81.08	76.92	78.95	65.22	35 GB
YOLO	IXI+150	68	8	10	92.69	79.42	**84.38**	89.47	87.18	**88.31**	**79.07**	39 MB

operating characteristics, none consistently identified FP SSS segmentations. InternVL3.5-14B adopted a conservative strategy, producing relatively few positive predictions, resulting in a macro precision of 52.71% but a macro recall of only 23.13%. Qwen2.5-VL-72B provided the best overall trade-off among the prompted models, reaching a micro F1 score of 44.33% and a J_{det} of 28.48%. In contrast, Gemma 3-31B detected nearly all positive instances, achieving a macro recall of 95.61%. However, this high sensitivity was accompanied by 191 FP predictions, leading to a macro precision of 27.40% and a J_{det} of 27.14%. Using only the original annotated projections (*IXI+10*), YOLO already outperformed all prompted VLMs while requiring only 39 MB of parameters. The detector achieved a pooled micro precision of 90.63% and a micro F1 score of 52.73%, exceeding the best prompted model by 8.40 percentage points (micro F1).

Task-specific training with projection-based view synthesis (*IXI+150*) yielded the best overall performance. Compared with the LoRA-adapted InternVL3.5-14B model, our approach improved the micro F1 score by 9.36 percentage points while requiring only 39 MB of parameters instead of 35 GB. These results demonstrate that the proposed task-specific training strategy is more effective for this quality assessment task than adapting substantially larger VLMs.

Effect of Projection-Based View Synthesis. To further investigate the effect of projection-based view synthesis, we analyze its impact on each SSS region. The class-specific results in Table 2 show that projection-based view synthesis consistently improved performance across all three SSS regions. For the anterior SSS, view synthesis increased recall from 36.84% to 86.84% while maintaining a high precision of 89.19%. Recall for the posterior region increased from 42.86% to 91.43%, resulting in an F1 score of 90.14%. The largest improvement was observed for the central SSS. While the detector trained without view synthesis failed to detect any of the five positive instances, the synthesized-view model correctly detected three without producing a false-positive central prediction. Nonetheless, as the sample size (n=5) is rather small, further experiments should be carried out to confirm the claim.

Figure 2 further illustrates the effect of view synthesis through the precision–recall (PR) and receiver operating characteristic (ROC) curves for the anterior and posterior SSS regions. Leftmost corresponds to the detector trained with the

Table 2: Class-specific performance on TubeTK. The test set contained 38 anterior, 35 posterior, and 5 central positive instances. P, R, and F1 denote precision, recall, and F1 score, respectively. All values are percentages.

		Anterior ($n = 38$)			Posterior ($n = 35$)			Central ($n = 5$)		
Method	Setting	P	R	F1	P	R	F1	P	R	F1
InternVL3.5-14B	2-shot	69.23	23.68	35.29	66.67	5.71	10.53	22.22	40.00	28.57
Qwen2.5-VL-72B	2-shot	55.26	55.26	55.26	48.57	48.57	48.57	11.63	100.00	20.83
Gemma 3-31B	2-shot	40.24	86.84	54.99	36.08	100.00	53.03	5.88	100.00	11.11
YOLO	IXI+10	93.33	36.84	52.83	88.24	42.86	57.69	0.00	0.00	0.00
InternVL3.5-14B LoRA	IXI+150	71.05	71.05	71.05	93.55	82.86	87.88	80.00	80.00	80.00
YOLO	IXI+150	89.19	86.84	88.00	88.89	91.43	90.14	100.00	60.00	75.00

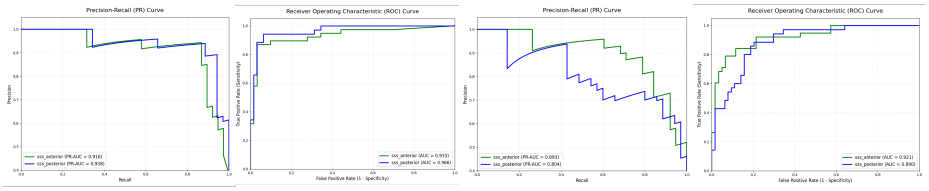

Fig. 2: PR and ROC curves for the anterior and posterior SSS regions. Left: detector trained with projection-based view synthesis. Right: detector trained on the original annotated projections only.

proposed view synthesis, whereas the rightmost shows the detector trained using only the original annotated projections. The improved curves indicate that the synthesized projections enhance the detector's discriminative performance across a broad range of operating thresholds, demonstrating that the observed gains are not specific to a single decision threshold.

These findings suggest that the synthesized projections introduced meaningful anatomical variability rather than merely increasing the number of nearly identical training images. In particular, variations in projection angle and slab selection exposed the detector to different appearances of the same vascular regions, improving its ability to generalize to unseen subjects.

6 Conclusion

We presented a canonical multi-view framework for quality assessment of cerebrovascular segmentations that reformulates localized three-dimensional quality assessment as a detection task in canonical two-dimensional projection views. By combining lightweight object detection with projection-based view synthesis, the proposed approach achieved substantially higher performance than prompted and LoRA-adapted VLMs while requiring only a fraction of their computational resources. Our work shows that localized quality assessment of complex vascu-

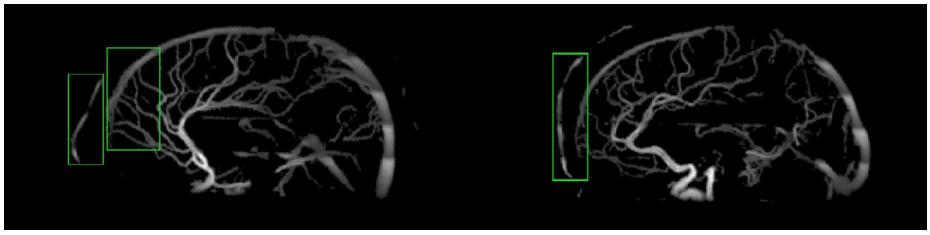

Fig. 3: Spurious skull segmentations detected as FP SSS. Left: jointly with the SSS. Right: detected as the SSS.

lar structures can be addressed through anatomically standardized projection views, avoiding the need for direct reasoning over the full 3D vascular tree.

The current framework has been demonstrated on the detection of erroneous SSS segmentations. While its extension to other localized cerebrovascular QC tasks and segmentation protocols appears straightforward, this would require extending the training data to include labels for additional structures, both vascular and non-vascular, a broader range of segmentation errors, and more fine-grained bounding boxes that better characterize the detected error type (Fig. 3).

Acknowledgments. This work is funded by the European Union (ERC CARAVEL, 101171357). J.C. is supported through a seed grant from the Royal College of Radiologists. We acknowledge computational resources provided by GENCI-IDRIS on Jean Zay A100 partition (2026-A0200617550).

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (Feb 2025). https://doi.org/10.48550/ARXIV.2502.13923
2. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
3. Cosarinsky, M., Billot, R., Mansilla, L., Jimenez, G., Gaggion, N., Fu, G., Tirer, T., Ferrante, E.: ConfIC-RCA: Statistically Grounded Efficient Estimation of Segmentation Quality. IEEE Transactions on Medical Imaging (2026). https://doi.org/10.1109/TMI.2026.3679618
4. Falcetta, D., Canas, L.S., Suppa, L., Pentassuglia, M., Cleary, J., Modat, M., Ourselin, S., Zuluaga, M.A.: An automated framework for large-scale graph-based cerebrovascular analysis. In: 2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2026)

5. Falcetta, D., Marciano, V., Yang, K., Cleary, J., Legris, L., Rizzaro, M.D., Pitsiorlas, I., Chaptoukaev, H., Lemasson, B., Menze, B., Zuluaga, M.A.: Vesselverse: A dataset and collaborative framework for vessel annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 655–665 (2025)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (Jun 2021). https://doi.org/10.48550/ARXIV.2106.09685
7. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements (Oct 2024). https://doi.org/10.48550/ARXIV.2410.17725
8. Köhler, P., Fadugba, J., Berens, P., Koch, L.M.: Efficiently correcting patch-based segmentation errors to control image-level performance in retinal images. In: Burgos, N., Petitjean, C., Vakalopoulou, M., Christodoulidis, S., Coupe, P., Delingette, H., Lartizien, C., Mateus, D. (eds.) Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 250, pp. 841–856. PMLR (03–05 Jul 2024), https://proceedings.mlr.press/v250/kohler24a.html
9. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023)
10. Marcianò, V., Chaptoukaev, H., Fernandez, V., Cardoso, M.J., Ourselin, S., Antonelli, M., Zuluaga, M.A.: Diffusion-based quality control of medical image segmentations across organs. arXiv preprint arXiv:2511.09588 (2025)
11. Xu, H., Nie, Y., Wang, H., Chen, Y., Li, W., Ning, J., Liu, L., Wang, H., Zhu, L., Liu, J., et al.: Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 391–401 (2025)