Pelvis-SigLIP: Benchmarking Vision-Language Models for Female Pelvic MRI Series Retrieval across Zero-Shot, Linear-Probe, and Fine-Tuning

Maximilian Lindholz[1,2], Richard Ruppel[1], Charlie Alexander Hamm[1,2], Anika Knupfer[3], Semil Eminovic[1], Robin Schmidt[1], Anna-Maria Haack[1], Yasmin El-Nahry[5], Carolina Dominguez Aleixo[1], Jana Hutter[3,4], Sylvia Mechsner[5], and Tobias Penzkofer[1,2]

[1] Department of Radiology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
[2] Berlin Institute of Health (BIH), Berlin, Germany
[3] Institut für Informationsverarbeitung, Leibniz University Hannover, Hannover, Germany
[4] King's College London, London, United Kingdom
[5] Department of Obstetrics, Charité – Universitätsmedizin Berlin, Berlin, Germany
maximilian.lindholz@charite.de

Abstract. Comparing a patient's MRI series across visits, or assembling a cohort from a large multi-site, multi-vendor database, both depend on the same step: retrieving the right series. Automating this is difficult because series names are inconsistent across scanner vendors, and often differ between sites using the same vendor. Vision-language models (VLMs) are a natural fit, embedding series thumbnails into a shared image-text space that can be queried with prompts such as "T2-weighted sagittal MRI of the female pelvis". We benchmark three medical VLMs (PubMedCLIP, BiomedCLIP, MedSigLIP) and one state-of-the-art general VLM (SigLIP-2) on 2,702 female pelvic MRI series from five cohorts, across three tiers: zero-shot classification, linear probing, and full fine-tuning. Zero-shot fails: every model scores 14.5 to 29.4% sequence accuracy, below the 36.2% majority-class baseline. This gap is not bridged by medical pretraining or by general-purpose scale. Fine-tuning closes it: all four models reach $\geq$93% sequence accuracy after end-to-end training. We release two fine-tuned SigLIP-based checkpoints as **Pelvis-SigLIP**: a fine-tuned MedSigLIP reaching 95.1% sequence accuracy, and a fine-tuned SigLIP-2 reaching 90.5% orientation accuracy on a held-out patient test set. Both are released for clinical research use.

Keywords: Vision-language models · MRI series retrieval · Female pelvic MRI · Zero-shot classification · Medical foundation models

1 Introduction

In female pelvic MRI, whether for cancer staging or for benign conditions such as endometriosis, radiologists compare examinations with prior studies, often

acquired on different scanners over several years. Different sequences carry complementary information, for instance T2-weighted imaging depicts the zonal anatomy of the uterus, while diffusion-weighted imaging (DWI) shows restricted diffusion [9, 13]. The practical obstacle is that the same sequence is often named differently across vendors: a T2-weighted sagittal series may appear as "T2W SAG" on GE, and "t2_tse_sag" on Siemens, sometimes with a corrupted orientation tag after a PACS migration. Series-description metadata is well known to be heterogeneous and unreliable across vendors and sites [1, 3].

Metadata-independent retrieval from the image itself solves this and comparable problems at three clinical scales: **(i)** temporal tracking, which means retrieving all T2 sagittals of a patient across visits, regardless of scanner vendor; **(ii)** research cohorts: querying a 10,000-patient database for all T2 sagittals without relying on inconsistent series names; **(iii)** acquisition quality control: verifying the correct sequences and orientations were acquired. VLMs are well suited to this task: by jointly embedding images and text [11, 12], they could let a radiologist query "T2 sagittal female pelvis" and retrieve matching series from any PACS, regardless of vendor. The natural input is the *series thumbnail*, the small preview image PACS workstations already display for each series and that clinicians use to identify sequences at a glance. Crucially, this requires no labelled examples, so VLMs are in principle deployable without annotation effort [11, 12].

Medical VLMs such as PubMedCLIP [2], BiomedCLIP [15] and MedSigLIP [12] are trained on large, partly non-public datasets that include some female pelvic images. In practice, however, female pelvic MRI is underrepresented in medical AI datasets and benchmarks [4]. This raises the question: does the models' general medical knowledge transfer to this domain, or is fine-tuning on pelvic-specific data necessary?

We evaluate three medical VLMs (PubMedCLIP, BiomedCLIP, MedSigLIP) and one general-domain baseline (SigLIP-2, not medically fine-tuned) on 2,702 female pelvic MRI series from five cohorts and three scanner vendors, across three evaluation tiers:

1. **Zero-shot**: classify series thumbnails using text prompts only, the zero-label deployment scenario.
2. **Linear Probe (LP)**: train a logistic classifier on frozen embeddings.
3. **Full fine-tuning (FT)**: end-to-end encoder fine-tuning, the ceiling performance.

Contributions. (1) To our knowledge, the first systematic VLM benchmark for female pelvic MRI series retrieval across five cohorts and three scanner vendors. **(2)** Zero-shot evaluation using PACS series thumbnails, the same visual cue clinicians use, revealing a domain gap for female pelvic MRI even for medical VLMs. **(3) Pelvis-SigLIP**: two released SigLIP-based checkpoints fine-tuned for female pelvic MRI series identification — MedSigLIP (95.1% sequence) and SigLIP-2 (90.5% orientation) — ready to deploy as image classifiers.

2 Related Work

Medical vision-language models. CLIP [11] showed that image-text contrastive pretraining enables zero-shot classification, and several works adapt this recipe to medical imaging: PubMedCLIP [2], BiomedCLIP [15], and MedSigLIP [12] differ mainly in the scale and source of their medical image-text corpora. SigLIP-2 [14] is a state-of-the-art general VLM, included to test whether web-scale pretraining alone can substitute for medical supervision. Architectural details of all four evaluated models are given in Section 3.

MRI series identification. Automatic sequence and orientation labeling has been approached through methods ranging from a random forest classifier [3] to deep-learning based approaches [1]. To our knowledge, VLM-based cross-vendor series identification remains unaddressed, particularly for female pelvic imaging.

3 Dataset and Models

Data. We aggregate **2,702 MRI series** from **534 patients** across five publicly accessible female pelvic MRI cohorts (CPTAC-UCEC [8], CCTH [7], TCGA-CESC [6], UT-EndoMRI [5], UMD [10]), spanning Siemens, GE, and Philips scanners at 1.5 T and 3.0 T. Each DICOM series is represented by a single 256×256 series thumbnail (central slice). Weak labels were generated by regex rules on DICOM series descriptions and iteratively refined under review of the corresponding images, by consensus of a medical-imaging computer scientist (8 years' experience) and a third-year radiology resident, until all labels matched the visual appearance of the series.

Two classification targets: **sequence type** (8 classes: T1, T1FS, T2, T2FS, DCE, DWI, ADC, other) and **view orientation** (4 classes: axial, sagittal, coronal, oblique). This grouping reflects a simplified clinical workflow and is not the only viable taxonomy; alternative schemes could, for instance, further separate contrast-enhanced from non-contrast acquisitions. Per-cohort statistics are shown in Table 1. Of the 2,702 selected series, the sequence types comprise 979 T2, 498 T1FS, 456 other, 319 T1, 117 DCE, 112 ADC, 112 T2FS, and 109 DWI, acquired in axial (1,242), sagittal (818), oblique (355), and coronal (287) orientations.

Vision-language models (main evaluation). We evaluate four VLMs spanning medical and general pretraining. **PubMedCLIP** [2] is a CLIP model (ViT-B/32, 512-dim) trained on PubMed figure-caption pairs, while **BiomedCLIP** [15] scales this idea to the 15-million-pair PMC corpus, pairing a ViT-B/16 encoder with PubMedBERT (512-dim). **MedSigLIP** [12] instead fine-tunes SigLIP-SO400M on medical image-text data (1152-dim), and **SigLIP-2** [14], our general-domain baseline, is a ViT-L/16-384 trained with a sigmoid contrastive loss (1024-dim). We L2-normalise all embeddings.

Table 1. Per-cohort dataset statistics

Dataset	Patients	Series	Vendors	Field
CCTH [7][a]	23	523	GE, Philips, Siemens	1.5 T
CPTAC-UCEC [8][a]	36	1,056	GE, Siemens	1.5 T
TCGA-CESC [6][a]	54	488	GE, Philips, Siemens	1.5/3.0 T
UMD [10]	300	300	Philips	1.5/3.0 T
UT-EndoMRI [5][b]	121	335	GE, Philips, Siemens	1.5/3.0 T
Total	**534**	**2,702**		

[a] Diagnostic MRI subsample only; full cohort contains additional non-diagnostic or non-MRI series. [b] Two sub-cohorts before exclusions: D1 (51 subjects, Memorial Hermann Hospital System) and D2 (82 subjects, Texas Children's Hospital Pavilion for Women). Subjects containing only segmentation labels without corresponding MRI were excluded, leaving 121 subjects.

4 Methods

We instantiate series retrieval as closed-set classification: a query prompt (zero-shot) or class prototype ranks all series in the database by cosine similarity, and top-1 accuracy is the precision of that ranking restricted to the label set.

Zero-shot classification. For each VLM, we classify each series thumbnail by cosine similarity between its L2-normalised image embedding and the mean text embedding of three prompt templates per class (e.g., "T2-weighted MRI of the female pelvis", "T2w pelvic MRI", "T2 MRI pelvic scan"). Prompts are encoded with each model's paired text encoder.

Linear Probe. L2-regularised logistic regression on frozen embeddings evaluated with 5-fold patient-level cross-validation (StratifiedGroupKFold on patient ID, outer loop); within each fold, $C \in \{0.01, \ldots, 100\}$ is selected via a nested 5-fold patient-level CV on the training portion, preventing leakage. We report mean $\pm$ std over outer folds.

Full fine-tuning. We fine-tune each VLM's vision encoder end-to-end for image classification (the text encoder is unused): a Xavier-initialised linear head ($\mathrm{Linear}(d_{\mathrm{emb}}, C)$) is appended and trained jointly with all encoder weights, with images resized to each model's native resolution. Sequence and orientation are trained as separate models, so no single backbone is updated for both objectives. *Data split:* Patients are split into a held-out 20 % test set and an 80 % train+val pool (GroupShuffleSplit on patient ID); the pool is divided into 5 StratifiedGroupKFold folds (patient-level, $\approx$64/16 % train/val). The best checkpoint per fold is evaluated on the fixed test set; we report mean $\pm$ std over folds. *Optimiser:* AdamW, differential learning rates ($\eta_{\mathrm{bb}}{=}1\times10^{-5}$, $\eta_{\mathrm{hd}}{=}1\times10^{-3}$), weight decay 0.01, cosine annealing, 3-epoch warmup; batch size 8 (MedSigLIP, SigLIP-2) /

32 (PubMedCLIP, BiomedCLIP). *Training:* Up to 50 epochs, early stopping (patience 10) on validation accuracy. Augmentation: horizontal flip (p=0.5), vertical flip (p=0.3), ±10° affine rotation.

5 Results

Table 2 shows results across all three evaluation tiers for all four VLMs.

5.1 Zero-Shot: Below the Majority-Class Baseline on Sequence

Despite being trained on medical image-text data, all four VLMs fail at zero-shot sequence classification. PubMedCLIP achieves 23.2%, BiomedCLIP 28.8%, MedSigLIP 29.4%, and SigLIP-2 14.5% (Table 2, ZS columns), at or barely above the 12.5% chance level for eight classes. Because the cohort is imbalanced, the more informative reference is the majority-class baseline: predicting T2 for every series already yields 36.2%, which no model reaches, so zero-shot sequence classification is worse than a trivial constant classifier. Figure 1 illustrates the zero-shot procedure: the model is given a set of textual class descriptions, and the class whose text embedding has the highest similarity to the image embedding is taken as the prediction. Orientation performance is higher but less uniform and overall still not sufficient for clinical deployment: SigLIP-2 (71.2%) and BiomedCLIP (66.4%) clear the 46.0% all-axial majority-class baseline, PubMedCLIP (50.7%) only modestly, and MedSigLIP (35.3%) falls below it.

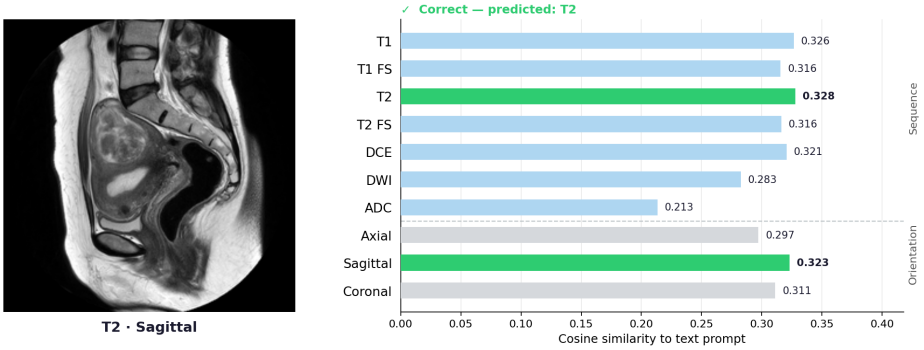

Fig. 1. Successful zero-shot text classification examples with PubMedCLIP, showing a text prompt and its corresponding PACS series thumbnail. Text prompts are simplified, and the "other/oblique" class is omitted, for clarity of display.

5.2 Linear Probing: Rapid Domain Adaptation

Linear probing, which is fast and memory-efficient because the backbone stays frozen, already recovers most of the lost accuracy: MedSigLIP rises from 29.4% to $92.4 \pm 0.9\%$ sequence accuracy. PubMedCLIP and SigLIP-2 reach $90.4 \pm 1.3\%$ and $89.4 \pm 1.3\%$, while BiomedCLIP reaches $83.3 \pm 1.5\%$, confirming that frozen VLM embeddings already encode substantial sequence-discriminative structure without any task-specific tuning. Part of this initial gain reflects the probe capturing the cohort's class priors, which the zero-shot prompts cannot access.

5.3 Full Fine-Tuning: Pelvis-SigLIP

End-to-end fine-tuning produces the best sequence accuracy across all models (Table 2, FT columns). On orientation the gain is smaller, and for MedSigLIP linear probing is marginally better. Our released **Pelvis-SigLIP** checkpoints lead on the two targets: fine-tuned MedSigLIP achieves the highest sequence accuracy ($95.1 \pm 0.8\%$), while fine-tuned SigLIP-2 leads on orientation ($90.5 \pm 0.8\%$) and is close on sequence ($95.0 \pm 1.1\%$). PubMedCLIP ($93.8 \pm 0.7\%$) and BiomedCLIP ($93.8 \pm 1.3\%$) achieve somewhat lower sequence accuracy. Notably, the general-domain SigLIP-2 is competitive with the medical VLMs after fine-tuning, indicating that for this task the quality of the base visual representation matters as much as medical pretraining. LP and FT use different protocols (5-fold patient-level CV over the full cohort vs. 5-fold CV on an 80% pool evaluated on a fixed held-out test set), so their standard deviations carry different sources of variability and the two columns are not paired; we therefore do not test LP–FT differences for significance.

Table 2. Evaluation ladder for four VLMs. **Bold** = best per column; ZS = zero-shot (text prompts, no labels); LP = Linear Probe (5-fold patient-level CV, nested C selection, mean ± std); FT = full fine-tuning (5-fold CV on 80%, held-out 20% test, mean ± std). LP and FT protocols differ and are therefore not paired. † Released as part of **Pelvis-SigLIP** (fine-tuned MedSigLIP and SigLIP-2).

	Sequence Type (8-class)			*View Orientation (4-class)*		
Model	ZS↑	LP↑	FT↑	ZS↑	LP↑	FT↑
PubMedCLIP	0.232	0.904±.013	0.938±.007	0.507	0.881±.015	0.890±.008
BiomedCLIP	0.288	0.833±.015	0.938±.013	0.664	0.862±.012	0.889±.005
MedSigLIP†	**0.294**	**0.924±.009**	**0.951±.008**	0.353	**0.893±.019**	0.891±.005
SigLIP-2†	0.145	0.894±.013	0.950±.011	**0.712**	0.882±.015	**0.905±.008**

Majority-class baseline: sequence = 36.2% (all T2), orientation = 46.0% (all axial).

5.4 Embedding Space Before and After Fine-Tuning

Figure 2 shows UMAP projections of MedSigLIP embeddings for sequence classification, before and after full fine-tuning. Embeddings are the L2-normalised

pooled image representations from the vision encoder, taken before the classification head in both cases. Before fine-tuning, the sequence types are only weakly separated, consistent with near-chance zero-shot accuracy. After fine-tuning, the same 2,702 series thumbnails form well-separated clusters matching the sequence classes, consistent with MedSigLIP's best-in-class fine-tuned sequence accuracy (95.1%).

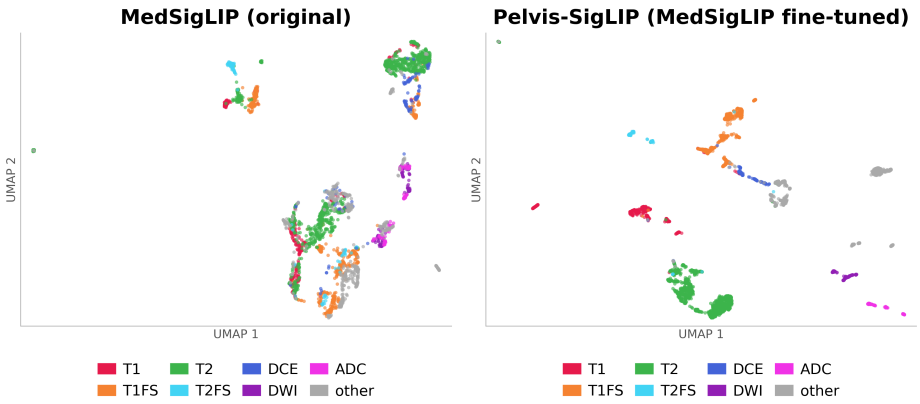

Fig. 2. UMAP of MedSigLIP embeddings from 2,702 female pelvic MRI series, **before** and **after** full fine-tuning for sequence classification.

6 Discussion and Conclusion

We benchmarked four VLMs for metadata-independent MRI series retrieval on 2,702 female pelvic series thumbnails across five cohorts. Zero-shot classification fails for all models (14.5 to 29.4% sequence accuracy, below the 36.2% majority-class baseline), including the three explicitly designed for medical imaging, likely reflecting the underrepresentation of female pelvic MRI in medical pretraining corpora [4]. The pretrained features nonetheless encode useful structure: linear probing on frozen embeddings recovers most of the performance at negligible compute, so task-specific adaptation rather than domain-specific pretraining is the bottleneck. This makes linear probing the pragmatic option for a new site adapting with limited resources; full fine-tuning adds a further 2.7 to 10.5 points on sequence (all four models $\geq$93%) but yields a task-specific encoder that no longer serves as a general-purpose foundation model. We release two fine-tuned SigLIP-based checkpoints as **Pelvis-SigLIP**: MedSigLIP, highest on sequence (95.1%), and SigLIP-2, highest on orientation (90.5%), both on a held-out patient test set. Two limitations follow from our design. First, we report classification accuracy rather than ranked-retrieval metrics (precision@k, mAP) over a full database; validating the ranking directly on a PACS-scale index is the natural

next step. Second, performance is summarised by overall accuracy, without per-class error analysis. More broadly, this work is a proof of concept for VLM-based MRI series identification that can be extended to more clinically tailored label sets beyond the generic taxonomy used here.

Acknowledgments. Code and model weights are available at: https://github.com/MaximilianLindholz/Pelvis-SigLIP. During the preparation of this work, the authors used Claude (Anthropic) to improve the readability and language of the manuscript and to assist with code development. All content was subsequently reviewed and edited by the authors, who take full responsibility for the content of the publication. Data used in this publication were generated by the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). Furthermore, the results shown here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Disclosure of Interests. M.L. receives funding from the Berlin Institute of Health (Junior Digital Clinician Scientist Grant). C.A.H. receives funding from the Berlin Institute of Health (Digital Clinician Scientist Grant) and Siemens Healthineers, previously received seed funding from the University Medicine Greifswald and speaker fees from Siemens Healthineers, Bayer AG, Koelis AG, and Bracco Imaging Deutschland outside the submitted work, and is on the advisory board of Lucida Medical. T.P. receives funding from the Berlin Institute of Health (Advanced Clinician Scientist Grant, Platform Grant), the German Federal Ministry of Research, Technology and Space (BMFTR, formerly BMBF), including the Network of University Medicine 3.0 (NUM 3.0), Grant No. 01KX2524, Project RACOON, which partially supported this publication, and the German Research Foundation (DFG, SFB 1340/2). T.P. declares research agreements with AGO, Astra Zeneca, Roche, Siemens Healthineers, and others (no personal payments); fees for a book translation (Elsevier B.V.); and fees for speaking engagements (Bayer Healthcare). A.K and J.H receive funding from CAIMed (Lower Saxony Center for Artificial Intelligence and Causal Methods in Medicine) [ZN4257]. The remaining authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baumgärtner, G.L., Hamm, C.A., Schulze-Weddige, S., Ruppel, R., Beetz, N.L., Rudolph, M., Dräger, F., Froböse, K.P., Posch, H., Lenk, J., Bießmann, F., Penzkofer, T.: Metadata-independent classification of MRI sequences using convolutional neural networks: Successful application to prostate MRI. European Journal of Radiology **166**, 110964 (2023). https://doi.org/10.1016/j.ejrad.2023.110964
2. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1181–1193 (2023)
3. Gauriau, R., Bridge, C., Chen, L., Kitamura, F., Tenenholtz, N.A., Kirsch, J.E., Andriole, K.P., Michalski, M.H., Bizzo, B.C.: Using dicom metadata for radiological image series categorization: a feasibility study on large clinical brain mri datasets. Journal of digital imaging **33**(3), 747–762 (2020)

4. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)
5. Liang, X., Alpuing Radilla, L.A., Khalaj, K., Dawoodally, H., Mokashi, C., Guan, X., Roberts, K.E., Sheth, S.A., Tammisetti, V.S., Giancardo, L.: A multi-modal pelvic mri dataset for deep learning-based pelvic organ segmentation in endometriosis. Scientific Data **12**(1), 1292 (2025)
6. Lucchesi, F.R., Aredes, N.D.: The cancer genome atlas cervical squamous cell carcinoma and endocervical adenocarcinoma collection (TCGA-CESC) (version 3) [data set] (2016). https://doi.org/10.7937/K9/TCIA.2016.SQ4M8YP4
7. Mayr, N., Yuh, W.T.C., Bowen, S., Harkenrider, M., Knopp, M.V., Lee, E.Y.P., Leung, E., Lo, S.S., Small Jr., W., Wolfson, A.H.: Cervical cancer – tumor heterogeneity: Serial functional and molecular imaging across the radiation therapy course in advanced cervical cancer (version 1) [data set] (2023). https://doi.org/10.7937/ERZ5-QZ59
8. National Cancer Institute CPTAC: CPTAC-UCEC: Clinical proteomic tumor analysis consortium — uterine corpus endometrial carcinoma. The Cancer Imaging Archive (2016). https://doi.org/10.7937/K9/TCIA.2016.GKJ0ZWAC
9. Nougaret, S., Horta, M., Sala, E., Lakhman, Y., Thomassin-Naggara, I., Kido, A., Masselli, G., Bharwani, N., Sadowski, E., Ertmer, A., et al.: Endometrial cancer mri staging: updated guidelines of the european society of urogenital radiology. European radiology **29**(2), 792–805 (2019)
10. Pan, H., Chen, M., Bai, W., Li, B., Zhao, X., Zhang, M., Zhang, D., Li, Y., Wang, H., Geng, H., et al.: Large-scale uterine myoma mri dataset covering all figo types with pixel-level annotations. Scientific Data **11**(1), 410 (2024)
11. 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)
12. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
13. Thomassin-Naggara, I., Dolciami, M., Chamie, L.P., Guerra, A., Bharwani, N., Freeman, S., Rousset, P., Manganaro, L.: Esur consensus mri for endometriosis: indications, reporting, and classifications. European Radiology **35**(11), 7260–7268 (2025)
14. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
15. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)