Proposal for a Swiss AI Initiative Large Project

Project Name

Beyond Signals and Structure: Enabling Generalizable Multimodal Cardiopulmonary AI

Scientific Lead

- Prof. Dr. Julia Vogt, ETH Zurich, julia.vogt@inf.ethz.ch

Co-applicants

- Prof. Dr. Ece Özkan Elsen, University of Basel, ece.oezkanelsen@unibas.ch
- Dr. Thomas Sutter, ETH Zurich, thomas.sutter@inf.ethz.ch

Partners

- Prof. Dr. André Euler, Kantonsspital Baden, andre.euler@ksb.ch
- Prof. Dr. Christian Matter, University Hospital Zurich, christian.matter@usz.ch
- Dr. Jedrzej Sarnecki, University Children's Hospital Basel (UKBB), jedrzej.sarnecki@ukbb.ch

Request Summary

Cost Dimension	Amount
Alps Compute Hours (GPU-h)	961,000 GPU-h
Alps Storage Space (TB)	10 TB
Number of files to store simultaneously	10 Mio
Swiss AI Funds for Personnel (CHF)	CHF 120'121
Matching Funds for Personnel (CHF)	CHF 120'121
Swiss AI Initiative Engineering (FTEs [%], number of months)	0.0

Proposal Description

a. Purpose of the Swiss AI Initiative Large Project

Cardiopulmonary diseases are multifaceted and cannot be fully understood from a single source of information. Clinicians routinely combine patient history, structured variables such as imaging, physiological signals, labs, medications, and expert reports for an integrated picture of disease presentation and progression. Current AI models, however, typically focus on single modalities or simplified vision-to-text mappings, which limits their generalizability and clinical relevance in practice.

To address this gap, we propose the first open-source multimodal cardiopulmonary AI model that integrates ultrasound (US) videos, electrocardiograms (ECG), chest X-rays (CXR), structured clinical variables, and unstructured documentation such as radiology and cardiology reports and clinical notes. Each modality contributes complementary insights: CXRs reveal structural abnormalities in the lungs and heart; US provides real-time morphology and function; ECG captures cardiac electrical activity over time; and clinical notes and reports embed expert interpretation and contextual data. By leveraging joint pretraining objectives, the model will better reflect how clinicians synthesize information across modalities for diagnosis and risk stratification. Jointly learning from these heterogeneous modalities will enable our model to support a broad range of downstream tasks, including improved diagnosis, disease progression modeling, and risk stratification. This multimodal integration is designed to improve robustness in the presence of missing or incomplete data and to enable holistic, modality-aware AI systems that align more closely with real-world clinical reasoning.

Building on our research on the development of a CXR vision language model (VLM) [1], the improvement of multimodal pretraining for CXRs [2, 3, 4], and a small Swiss AI Grant combining US and ECG [5, 6], this project will extend the current state-of-the-art in multimodal medical AI, not only on a methodological level but also in terms of computational scale. We assembled a multimodal cardiopulmonary dataset of 6.7 Mio datapoints and five different modalities. Our novel training objectives, specifically designed for the multimodal nature of healthcare data, are beneficial over simplified bimodal image-text pretraining methods that form the basis for general-purpose VLMs. We envision the integration of the Swiss LLM Apertus [7, 8] for multimodal finetuning to build a Swiss-made cardiopulmonary AI system. The experience we gained in previous Swiss AI grants allows us to tackle the computational and engineering challenges involved in building a multimodal AI model that natively integrates the information from a diverse set of modalities.

Realizing such a large-scale project requires a cross-institutional team of AI, engineering, and medical experts. We built a team of AI and engineering experts from the groups of Prof. Vogt and Prof. Özkan Elsen, and medical experts from cardiology, radiology, and pediatric radiology.

We aim to learn unified cardiopulmonary representations that combine visual, physiological, textual, and clinical inputs and build the first natively multimodal cardiopulmonary foundation model that integrates all relevant modalities already during pretraining (see Figure 1). This approach has the potential to support more accurate diagnosis and risk stratification for complex cardiopulmonary conditions while improving adaptability to real-world clinical variability. Our broader objective is to develop generalizable, modality-aware AI systems that can be scaled across diverse healthcare settings and tailored towards the clinical needs.

b. Current Situation of the Thematic Area

Pretraining. Self-supervised pretraining has become the default paradigm for label-efficient learning. Contrastive methods such as SimCLR [9] and MoCo [10], as well as masking-based approaches such as masked autoencoders [MAE, 11], learn transferable visual features from large unlabelled datasets. Large language models such as BERT [12] and GPT [13] enable zero-/few-shot behaviors in text. Their bimodal descendants, including CLIP [14] and ALIGN [15], align image–text spaces.

Pretraining for healthcare data. Medical approaches in healthcare AI for CXR, EHR, US, and ECG leverage uni- or bimodal pretraining strategies. Ultrasound (US) models like SAMUS [16], UltraSam [17], USFM [18], and UltraFedFM [19] demonstrate large-scale segmentation and generalization, while echocardiography (ECHO) foundation models such as EchoFM [20, 21], and EchoPrime [22]

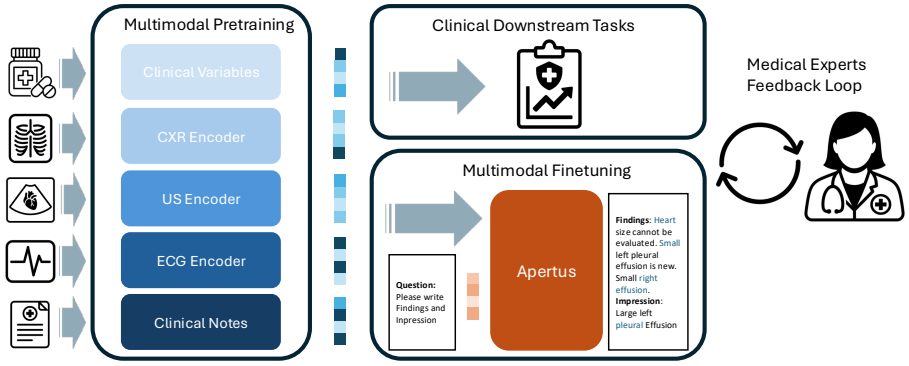

Figure 1: We incorporate visual, physiological, textual, and clinical inputs and build the first natively multimodal cardiopulmonary AI model that integrates all relevant modalities already during pretraining. In close collaboration with medical experts, we evaluate our multimodal encoders on diverse clinical downstream tasks and incorporate a multimodal finetuning pipeline for relevant tasks in diagnosing cardiopulmonary diseases.

expand video-based cardiac function assessment. ECG [23, 24, 25] and CXR [26, 27, 28] foundation models achieve strong diagnostic accuracy and transfer learning through large-scale pretraining. Healthcare vision–language models such as MedCLIP [29], EchoCLIP [30], BioViL [31], as well as generalist medical systems like GatorTron [32], Med-PaLM [33], and Med-Gemini [34].

Large-Scale multimodal pretraining. A more recent line of work has shifted toward truly large-scale multimodal pretraining that integrates many modalities simultaneously, often including vision, text, audio, video, tabular data, and biological signals. ImageBind[35], [36], and X-VILA [37] demonstrated that it is possible to align six modalities (images, text, audio, depth, thermal, IMU) into a single shared embedding space, even without all-to-all paired data. Multimodal pretraining requires not only scaling data and compute, but also developing strategies for modality alignment, missing modality handling, and shared semantic representation.

In the healthcare domain, where data is naturally multimodal, approaches that leverage the combined multimodal structure are especially important. However, biomedical foundation models remain in the bimodal pretraining setting [38]. Hence, there is a clear need for a multimodal cardiopulmonary foundation model that integrates and leverages the information coming from CXR, US, ECG, EHR, notes, and reports.

MIMIC-IV. The MIMIC "universe" provides interoperable modules spanning clinical tables, free-text notes, imaging, and physiological signals, enabling broad pretraining and rigorous ablations. MIMIC-IV offers linked hospital EHR [39] and deidentified notes [40]; MIMIC-IV-CXR adds over 377,000 chest radiographs with reports [41]; MIMIC-IV-ECG contributes roughly 800,000 12-lead diagnostic ECGs matched to cardiologist reports [42]; and MIMIC-IV-Echo provides over 500,000 echocardiography videos with associated structured measurements and reports [43]. These resources make MIMIC-IV a natural substrate for developing unified cardiopulmonary foundation models spanning clinical variables, notes, CXR, US, ECG, and radiology and cardiology reports. Unlike previous works, we will also tackle the large-scale multimodal extension of MIMIC-IV with other datasets.

Summary. Despite unimodal models across cardiopulmonary-relevant modalities, large-scale integration across imaging, ECG, ultrasound, structured EHR, and clinical text remains rare. However, bridging these modalities with scalable, unified pretraining will enable general-purpose diagnostic and prognostic models for cardiopulmonary diseases.

c. Activities

We propose five different work packages to achieve the goals and ideas outlined above that are summarized in Figure 2. We describe the individual work packages in more detail below.

WP1: Preparation of datasets. This work package builds on our existing, well-established unimodal data pipelines for US, ECGs, and CXRs, which already include preprocessing steps such as quality filtering, temporal alignment, normalization, and metadata extraction. In WP1.1, we will extend this

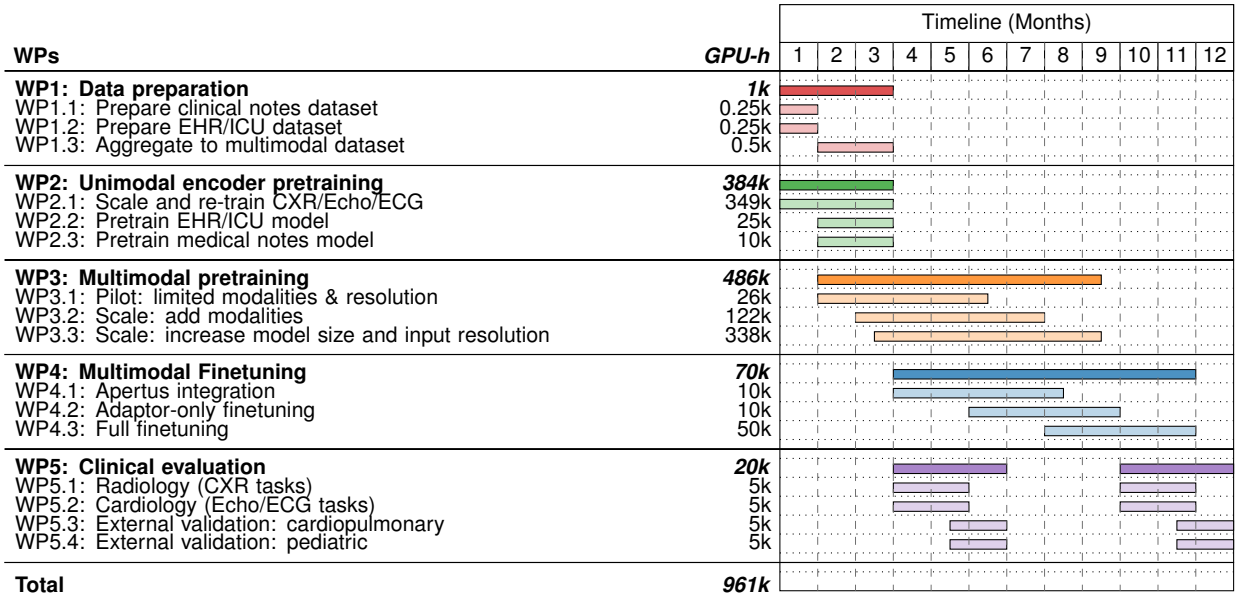

Figure 2: Overview of work packages including their computational demand and timeline.

framework to structured clinical variables by mapping them to standardized formats and addressing time-stamped missingness. In WP1.2, we will prepare medical notes and radiology/cardiology reports by sectioning and tokenizing free text for downstream modeling. In WP1.3, we will construct a multimodal integration pipeline to ensure consistent patient- and study-level alignment across all modalities, addressing challenges such as variable sampling rates, asynchronous acquisition, and modality-specific missing data. The deliverable of this work package will be a clean, harmonized, and temporally aligned multimodal dataset, ready joint multimodal learning in subsequent stages.

WP2: Unimodal encoder pretraining. To ensure a robust multimodal modeling pipeline, we adopt a staged approach that begins with unimodal encoder pretraining. This phase allows us to validate data quality, refine architecture choices, and establish strong baseline encoders before multimodal integration. In WP2.1, we will scale up and re-train existing CXR, Echo, and ECG encoders developed in our small grant, revisiting key hyperparameters and design decisions. In WP2.2, we will train unimodal encoders for structured EHR/ICU data, focusing on temporal modeling of clinical variables. In WP2.3, we will pretrain models for medical notes and reports, leveraging domain-specific transformers for tokenized free-text input. The deliverable of WP2 will be a set of well-performing unimodal encoders across all modalities, serving as reliable building blocks for multimodal pretraining and enabling direct comparison between unimodal and multimodal objectives in WP4.

WP3: Multimodal pretraining. Once the unimodal encoders are established, we proceed to joint multimodal training to learn cross-modal interactions and shared representations. The goal is to pretrain a large multimodal model that integrates all available clinical modalities. In WP3.1, we will begin with small-scale multimodal training that integrates a limited number of modalities at lower resolution to validate training objectives. In WP3.2, we will scale up the number of integrated modalities, extending beyond vision–text to include structured EHR and ICU data and unstructured notes. In WP3.3, we will scale up the spatial resolution of imaging and video modalities to improve fine-grained cardiopulmonary analysis. To adjust to the increased resolution, we will increase the model size to assess the effect of capacity on multimodal alignment. Across these steps, we will implement cross-modal learning objectives such as contrastive alignment, masked prediction, and modality dropout, strategies already validated in the first phase of the Swiss AI initiative. The deliverable of WP3 will be a unified multimodal representation space optimized to handle heterogeneous and partially missing clinical data, with ablation studies quantifying the contribution of each modality and benchmarking scalability on downstream diagnostic and prognostic tasks.

WP4: Multimodal finetuning Integrating our multimodal encoders with LLMs, such as the Swiss LLM Apertus [7, 8], offers a powerful framework for understanding and reasoning over cardiopulmonary data. WP4 is designed to enable a more flexible interaction via natural language, and support a more

diverse set of tasks, such as question answering over multimodal encodings of cardiopulmonary data. In WP4.1, we will integrate Apertus into our existing multimodal LLM finetuning pipeline. In WP4.2, we will test the quality of our multimodal encoders in combination with Apertus. For these first tests, we will restrict the finetuning to the adaptor networks that connect the multimodal encodings with the LLM. With a limited computational load, we can produce first results and gain understanding about the AI system simultaneously. In WP4.3, we will finetune all parameters, i.e., the multimodal encoders, the adaptors, and the LLM. We expect to boost the performance but also increase the computational load.

WP5: Evaluation of clinical relevance. To ensure that the developed multimodal models are not only technically sound but also clinically meaningful, we will conduct a comprehensive evaluation focusing on real-world medical relevance. This includes assessing the models' ability to support key diagnostic, prognostic, and decision-support tasks across cardiology and pulmonology, such as identifying structural heart abnormalities, detecting arrhythmias, or stratifying cardiopulmonary risk. Beyond conventional computer science–driven performance metrics (e.g., AUROC, F1), we will prioritize clinical interpretability, robustness across patient subgroups, and alignment with expert reasoning. Our medical collaborators will play a central role in this process by providing domain expertise to validate model outputs, design clinically grounded evaluation protocols, and interpret failure cases. This collaborative evaluation process will help ensure that the final models are trustworthy, useful in practice, aligned with clinical workflows, and relevant for clinical practice.

d. Team

The Medical Data Science (MDS) group, led by Prof. Dr. Julia Vogt, works at the intersection of fundamental machine learning research and clinical applications, with a particular focus on multimodal AI for healthcare. The proposed MDS team brings together complementary expertise in large-scale representation learning, multimodal integration, and clinical translation. Thomas Sutter, Andrea Agostini, Sonia Laguna, Samuel Ruiperez-Campillo, and Moritz Vandenhirtz contributed to the first phase of the Swiss AI initiative [1, 2]. Thomas Sutter [3, 5, 6, 44, 45, 46] developed novel multimodal pretraining strategies for chest X-rays and ultrasound, providing a strong methodological foundation for this project. Andrea Agostini [4, 2, 6, 3], in addition, provides the required engineering expertise to make the novel methods scale to the Alps infrastructure. Moritz Vandenhirtz [47, 48], Samuel Ruiperez-Campillo [49, 50, 51, 52, 53, 54], and Sonia Laguna [55, 56, 57, 58] extend this expertise with backgrounds in interpretable and temporal modeling. The Analytics and Informatics for Child Health (AICH) group, led by Prof. Dr. Ece Özkan Elsen [59, 60, 61], brings domain expertise in US imaging and clinical text modeling, with a focus on sequential and multi-view learning methods for real-time medical data. Sergio Muñoz-González [62] and Simon Böhi [63] add expertise in machine learning for physiological signals and clinical decision support. Sergio Muñoz-González also provides expertise in ML engineering. The project is further supported by close collaboration with clinical experts: Prof. Dr. Andre Euler [64, 65] and Dr. Jedrzej Sarnecki [66, 67] as radiologists, and Prof. Dr. Christian Matter [68, 69, 70] as cardiologist provide essential domain knowledge, annotated private datasets, and clinical evaluation of model outputs. The team combines cutting-edge machine learning research with direct clinical expertise, ensuring the development of robust and multimodal models that address both technical challenges and clinical needs.

e. Expected Outputs and Outcomes

Within the one-year funding period, the project will deliver concrete outputs in the form of models, benchmarks, and open-source tools. Specifically, we will (i) train and release multimodal foundation model weights based on MIMIC-IV modules using ECG, US, CXR, EHR, and clinical text data, distributed either openly or through controlled-access repositories (e.g., PhysioNet) depending on licensing; (ii) provide a complete open-source codebase covering data preprocessing, multimodal integration, training, and evaluation workflows; and (iii) establish benchmarks and metrics for cardiopulmonary multimodal learning tasks, ensuring standardized and reproducible evaluation. Where

direct data sharing is restricted, we will supply reproducibility protocols, configuration files, and synthetic examples to allow other groups to replicate our results. We will publish results at leading venues in machine learning and healthcare (e.g., NeurIPS, ICML, ICLR, MLHC, CHIL) and make pretrained models and evaluation pipelines easily accessible to the community.

f. Importance for Switzerland, Europe, and beyond

This project directly advances the goals of the Swiss AI Initiative by developing open, reproducible, and clinically relevant foundation models for cardiopulmonary disease, the leading cause of morbidity and mortality worldwide. By integrating ECG, US, chest radiography, EHR, and clinical text—modalities that current AI systems mostly treat in isolation—we address a critical gap in multimodal clinical AI. Demonstrating large-scale pretraining on openly accessible clinical data and releasing tools, benchmarks, and model weights will establish a platform that can be scaled to national and international datasets, positioning Switzerland at the forefront of trustworthy, high-impact healthcare AI. At the European level, the project complements ongoing work on AI regulation and responsible data use by providing transparent, open-source resources that set new standards for reproducibility. Globally, it contributes to the demand for generalizable, clinically grounded AI systems and ensures broad impact through openly available models and tools that support both high-resource and resource-limited healthcare settings.

g. Preliminary Activities and Feasibility Report

In the attached feasibility report, we back the number of requested hours using the job reporting tool and scaling properties of our pipeline. The attached feasibility report shows the sophistication of our pipeline, both in data management and the engineering of AI systems. In addition, the MDS group has been part of the Swiss AI initiative since its start in early 2024. We developed a multimodal assistant for CXR interpretation [RadVLM, 1] using the Alps cluster under the vertical health project [1], see Table 1 for a breakdown of the GPU hours consumed during RadVLM development. In addition, we developed a novel multimodal pretraining paradigm [3], which we evaluated on CXR data [4]. Using the Alps infrastructure, we could show the positive effects of our novel pretraining paradigm on CXR downstream task performance [2]. Please see Table 2 for details on the GPU hours used. During the first phase of the Swiss AI initiative, we gained the knowledge, experience, and skills needed to leverage the infrastructure provided by CSCS. We pretrained various multimodal methods at different model and dataset configurations, including dataset and model sizes, different combinations of input modalities, and learning tasks (see Appendix A for more details). Since early this summer, we have been extending the developed pretraining paradigm to datasets of US and ECG as part of a Swiss AI small grant, where we already see early promising results both on US [5] and ECG [6]. see also Table 3. In summary, the previous projects have equipped us well with the required tools and experience, both on the engineering level to leverage the Alps infrastructure as well as on the machine learning modeling side to develop this novel multimodal medical method.

h. Ethics and Regulatory Compliance

This project is fully committed to ethical, legal, and regulatory standards in clinical AI research. We will use only de-identified, publicly available datasets (e.g., MIMIC-IV and modules) that comply with HIPAA and PhysioNet policies, with all team members completing certified training in human subjects research. Data handling will follow strict safeguards to ensure confidentiality, integrity, and traceability, and all models, code, and benchmarks will be released under licenses consistent with original data-use agreements. Where direct data sharing is restricted, we will provide reproducibility protocols, documentation, and synthetic examples. Compliance with the Swiss Federal Act on Data Protection (FADP) and the EU GDPR will guarantee that no re-identification is possible and that outputs cannot be misused for discriminatory purposes, ensuring responsible and transparent AI development for healthcare.

References

[1] Nicolas Deperrois, Hidetoshi Matsuo, Samuel Ruipérez-Campillo, Moritz Vandenhirtz, Sonia Laguna, Alain Ryser, Koji Fujimoto, Mizuho Nishio, Thomas M. Sutter, Julia E. Vogt, Jonas Kluckert, Thomas Frauenfelder, Christian Blüthgen, Farhad Nooralahzadeh, and Michael Krauthammer. RadVLM: A multitask conversational vision-language model for radiology, 2025.

[2] Andrea Agostini, Sonia Laguna, Alain Ryser, Samuel Ruiperez-Campillo, Moritz Vandenhirtz, Nicolas Deperrois, Farhad Nooralahzadeh, Michael Krauthammer, Thomas M Sutter, and Julia E Vogt. Leveraging the structure of medical data for improved representation learning. *arXiv preprint arXiv:2507.02987*, 2025.

[3] Thomas Sutter, Yang Meng, Andrea Agostini, Daphné Chopard, Norbert Fortin, Julia Vogt, Babak Shahbaba, and Stephan Mandt. Unity by diversity: Improved representation learning for multimodal VAEs. *Advances in Neural Information Processing Systems*, 37:74262–74297, 2024.

[4] Andrea Agostini, Daphné Chopard, Yang Meng, Norbert Fortin, Babak Shahbaba, Stephan Mandt, Thomas M Sutter, and Julia E Vogt. Weakly-supervised multimodal learning on MIMIC-CXR. In *Machine Learning for Health (ML4H)*, 2024.

[5] Yves Stebler, Thomas M. Sutter, Ece Ozkan, and Julia E. Vogt. Temporal representation learning for ultrasound analysis using masked modeling, 2025.

[6] Lucas Erlacher, Andrea Agostini, Samuel Ruiperez-Campillo, Ece Ozkan, Thomas M. Sutter, and Julia E. Vogt. Swiss-beatsnet: A multilead masked autoencoder for chagas disease detection, 2025.

[7] ETH Zurich. A language model built for the public good, 2025. Accessed: 2025-08-22.

[8] Dongyang Fan, Vinko Sabolčec, Matin Ansaripour, Ayush Kumar Tarun, Martin Jaggi, Antoine Bosselut, and Imanol Schlag. Can performant llms be ethical? quantifying the impact of web crawling opt-outs, 2025.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2019.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2018.

[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

[16] Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting SAM: Towards end-to-end ultrasound image segmentation via auto prompting. page 24–34, 2024.

[17] Adrien Meyer, Aditya Murali, Didier Mutter, and Nicolas Padoy. UltraSam: A foundation model for ultrasound using large open-access segmentation datasets. *arXiv preprint arXiv:2411.16222*, 2024.

[18] Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, and Yi Guo. USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical Image Analysis*, 96:103202, 2024.

[19] Yuncheng Jiang, Chun-Mei Feng, Jinke Ren, Jun Wei, Zixun Zhang, Yiwen Hu, Yunbi Liu, Rui Sun, Xuemei Tang, Juan Du, Xiang Wan, Yong Xu, Bo Du, Xin Gao, Guangyu Wang, Shaohua Zhou, Shuguang Cui, Rick Siow Mong Goh, Yong Liu, and Zhen Li. Privacy-preserving federated foundation model for generalist ultrasound artificial intelligence. *arXiv preprint arXiv:2411.16380*, 2024.

[20] Sekeun Kim, Pengfei Jin amd Sifan Song, Cheng Chen, Yiwei Li, Hui Ren, Xiang Li, Tianming Liu, and Quanzheng Li. EchoFM: Foundation model for generalizable echocardiogram analysis. *arXiv:2410.23413*, 2024.

[21] Ziyang Zhang, Qinxin Wu, Sirui Ding, Xiaolong Wang, and Jiancheng Ye. Echo-Vision-FM: A pre-training and fine-tuning framework for echocardiogram video vision foundation model. October 2024.

[22] Milos Vukadinovic, Xiu Tang, Neal Yuan, Paul Cheng, Debiao Li, Susan Cheng, Bryan He, and David Ouyang. EchoPrime: A multi-video view-informed vision-language model for comprehensive echocardiography interpretation, 2024.

[23] Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2024.

[24] Yue Wang, Xu Cao, Yaojun Hu, Haochao Ying, James Matthew Rehg, Jimeng Sun, Jian Wu, and Jintai Chen. AnyECG: Foundational models for electrocardiogram analysis. *arXiv preprint arXiv:2411.17711*, 2024.

[25] Lucas Bickmann, Lucas Plagwitz, Antonius Büscher, Lars Eckardt, and Julian Varghese. ExChanGeAI: An end-to-end platform and efficient foundation model for electrocardiogram analysis and fine-tuning. *arXiv preprint arXiv:2503.13570*, 2025.

[26] Weijian Huang, Cheng Li, Hong-Yu Zhou, Hao Yang, Jiarun Liu, Yong Liang, Hairong Zheng, Shaoting Zhang, and Shanshan Wang. Enhancing representation in radiography-reports foundation model: a granular alignment algorithm using masked contrastive learning. *Nature Communications*, 15(1), September 2024.

[27] Zefan Yang, Xuanang Xu, Jiajin Zhang, Ge Wang, Mannudeep K. Kalra, and Pingkun Yan. Chest x-ray foundation model with global and local representations integration, 2025.

[28] Jingfeng Yao, Xinggang Wang, Yuehao Song, Huangxuan Zhao, Jun Ma, Yajie Chen, Wenyu Liu, and Bo Wang. Eva-x: A foundation model for general chest x-ray analysis with self-supervised learning, 2024.

[29] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022.

[30] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5):1481–1488, April 2024.

[31] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. *Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing*, page 1–21. Springer Nature Switzerland, 2022.

[32] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records, 2022.

[33] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, July 2023.

[34] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, Eric Wang, Ellery Wulczyn, Fayaz Jamil, Theo Guidroz, Chuck Lau, Siyuan Qiao, Yun Liu, Akshay Goel, Kendall Park, Arnav Agharwal, Nick George, Yang Wang, Ryutaro Tanno, David G. T. Barrett, Wei-Hung Weng, S. Sara Mahdavi, Khaled Saab, Tao Tu, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, Jorge Cuadros, Gregory Sorensen, Yossi Matias, Katherine Chou, Greg Corrado, Joelle Barral, Shravya Shetty, David Fleet, S. M. Ali Eslami, Daniel Tse, Shruthi Prabhakara, Cory McLean, Dave Steiner, Rory Pilgrim, Christopher Kelly, Shekoofeh Azizi, and Daniel Golden. Advancing multimodal medical capabilities of Gemini, 2024.

[35] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all, 2023.

[36] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2024.

[37] Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, and Hongxu Yin. X-vila: Cross-modality alignment for large language model, 2024.

[38] Yunhao Liu, Suyang Xi, Shiqi Liu, Hong Ding, Chicheng Jin, Chenxi Yang, Junjun He, and Yiqing Shen. Multimodal medical image binding via shared text embeddings, 2025.

[39] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), January 2023.

[40] Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv-note: Deidentified free-text clinical notes, 2023.

[41] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019.

[42] Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset, 2023.

[43] Brian Gow, Tom Pollard, Nathaniel Greenbaum, Benjamin Moody, Alistair Johnson, Elizabeth Herbst, Jonathan W Waks, Parastou Eslami, Ashish Chaudhari, Tanner Carbonati, Seth Berkowitz, Roger Mark, and Steven Horng. Mimic-iv-echo: Echocardiogram matched subset, 2023.

[44] Thomas M Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal ELBO. *arXiv preprint arXiv:2105.02470*, 2021.

[45] Thomas Sutter, Imant Daunhawer, and Julia Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. *Advances in neural information processing systems*, 33:6100–6110, 2020.

[46] Imant Daunhawer, Thomas M Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal vaes. In *International Conference on Learning Representations*.

[47] Sarah Hagmann, Venkat Ramakrishnan, Alexander Tamalunas, Marc Hofmann, Moritz Vandenhirtz, Silvan Vollmer, Jsmea Hug, Philipp Niggli, Antonio Nocito, Rahel A Kubik-Huch, et al. Two decades of active surveillance for prostate cancer in a single-center cohort: favorable outcomes after transurethral resection of the prostate. *Cancers*, 14(2):368, 2022.

[48] Moritz Vandenhirtz and Julia E Vogt. From pixels to perception: Interpretable predictions via instance-wise grouped feature selection. In *Forty-second International Conference on Machine Learning*.

[49] Francisco Castells*, Samuel Ruipérez-Campillo*, Izan Segarra, Raquel Cervigón, Rubén Casado-Arroyo, José Luis Merino, and José Millet. Performance assessment of electrode configurations for the estimation of omnipolar electrograms from high density arrays. *Computers in biology and medicine*, 154:106604, 2023.

[50] Maarten ZH Kolk, Samuel Ruipérez-Campillo, Laura Alvarez-Florez, Brototo Deb, Erik J Bekkers, Cornelis P Allaart, Anne-Lotte CJ Van Der Lingen, Paul Clopton, Ivana Išgum, Arthur AM Wilde, et al. Dynamic prediction of malignant ventricular arrhythmias using neural networks in patients with an implantable cardioverter-defibrillator. *Lancet eBioMedicine*, 99, 2024.

[51] Johanna B Tonko*, Samuel Ruipérez-Campillo*, Gema Cabero-Vidal, Eva Cabrera-Borrego, Caroline Roney, Juan Jiménez-Jáimez, José Millet, Francisco Castells, and Pier D Lambiase. Vector field heterogeneity as a novel omnipolar mapping metric for functional substrate characterization in scar-related ventricular tachycardias. *Heart Rhythm*, 2024.

[52] Maarten ZH Kolk, Samuel Ruipérez-Campillo, Cornelis P Allaart, Arthur AM Wilde, Reinoud E Knops, Sanjiv M Narayan, Fleur VY Tjong, and DEEP RISK investigators Raijmakers Femke D. 1 2 Van Der Lingen Anne-Lotte CJ 5 Götte Marco JW 5 Selder Jasper L. 5 Alvarez-Florez Laura 6 Išgum Ivana 6 Bekkers Erik J. 7. Multimodal explainable artificial intelligence identifies patients with non-ischaemic cardiomyopathy at risk of lethal ventricular arrhythmias. *Scientific Reports*, 14(1):14889, 2024.

[53] Samuel Ruipérez-Campillo, José Millet, and Francisco Castells. Classification of atrial tachycardia types using dimensional transforms of ecg signals and machine learning. In *2022 Computing in Cardiology (CinC)*, volume 498, pages 1–4. IEEE, 2022.

[54] Samuel Ruipérez-Campillo, Alain Ryser, Thomas M Sutter, Maarten ZH Kolk, Fleur VY Tjong, Sanjiv M Narayan, and Julia E Vogt. A denoising vae for intracardiac time series in ischemic cardiomyopathy. *12th international conference on learning representations - TS4H*.

[55] Karthik Gopinath, Andrew Hoopes, Daniel C Alexander, Steven E Arnold, Yael Balbastre, Benjamin Billot, Adrià Casamitjana, You Cheng, Russ Yue Zhi Chua, Brian L Edlow, et al. Synthetic data in generalizable, learning-based neuroimaging. *Imaging Neuroscience*, 2:1–22, 2024.

[56] Daphné Chopard, Sonia Laguna, Kieran Chin-Cheong, Annika Dietz, Anna Badura, Sven Wellmann, and Julia E Vogt. Automatic classification of general movements in newborns. *arXiv preprint arXiv:2411.09821*, 2024.

[57] Sonia Laguna, Riana Schleicher, Benjamin Billot, Pamela Schaefer, Brenna Mckaig, Joshua N Goldstein, Kevin N Sheth, Matthew S Rosen, W Taylor Kimberly, and Juan Eugenio Iglesias. Super-resolution of portable low-field mri in real scenarios: integration with denoising and domain adaptation. In *Medical Imaging with Deep Learning*, 2022.

[58] Emanuele Palumbo, Laura Manduchi, Sonia Laguna, Daphné Chopard, and Julia E Vogt. Deep generative clustering with multimodal diffusion variational autoencoders. In *The Twelfth International Conference on Learning Representations*, 2024.

[59] Ričards Marcinkevičs, Patricia Reis Wolfertstetter, Ugne Klimiene, Kieran Chin-Cheong, Alyssia Paschke, Julia Zerres, Markus Denzinger, David Niederberger, Sven Wellmann, Ece Ozkan, Christian Knorr, and Julia E. Vogt. Interpretable and intervenable ultrasonography-based machine learning models for pediatric appendicitis. *Medical Image Analysis*, 91:103042, January 2024.

[60] Hanna Ragnarsdottir, Ece Ozkan, Holger Michel, Kieran Chin-Cheong, Laura Manduchi, Sven Wellmann, and Julia E. Vogt. Deep learning based prediction of pulmonary hypertension in newborns using echocardiograms. *International Journal of Computer Vision*, 132(7):2567–2584, February 2024.

[61] Ece Ozkan, Thomas M. Sutter, Yurong Hu, Sebastian Balzer, and Julia E. Vogt. M(otion)-mode based prediction of ejection fraction using echocardiograms. page 307–320, 2024.

[62] Carolina Simó, Maricruz Mamani-Huanca, Oswaldo Hernández-Hernández, Álvaro Redondo-Río, Sergio Muñoz, and Virginia García-Cañas. Application of nanopore long-read sequencing and metabolomics in an in vitro dynamic intestinal digestion model: A genome-centric metatranscriptomic approach to investigating microbial tma and scfa metabolism. *Journal of Pharmaceutical and Biomedical Analysis*, 262:116896, September 2025.

[63] Böhi, Simon and Gashi, Shkurta. Large language models for wearable data analysis and interpretation. 2024.

[64] Darin P. Clark, Fides R. Schwartz, André Euler, Victor Mergen, Hatem Alkadhi, Daniele Marin, and Cristian T. Badea. Unsupervised learning of robust models for cardiac and photon-counting x-ray ct denoising. In Rebecca Fahrig, John M. Sabol, and Lifeng Yu, editors, *Medical Imaging 2023: Physics of Medical Imaging*, page 60. SPIE, April 2023.

[65] Dusan Pisarcik, Marc Kissling, Jakob Heimer, Monika Farkas, Cornelia Leo, Rahel A. Kubik-Huch, and André Euler. Artificial intelligence language models to translate professional radiology mammography reports into plain language – impact on interpretability and perception by patients. *Academic Radiology*, 32(9):4988–4996, September 2025.

[66] Agata Paszkowska, Jedrzej Sarnecki, Alicja Mirecka-Rola, Monika Kowalczyk-Domagała, Lukasz Mazurkiewicz, and Lidia Ziolkowska. Imaging features of pediatric left ventricular noncompaction cardiomyopathy in echocardiography and cardiovascular magnetic resonance. *Journal of Cardiovascular Development and Disease*, 9(3):77, March 2022.

[67] Jedrzej Sarnecki, Agata Paszkowska, Joanna Petryka-Mazurkiewicz, Agata Kubik, Janusz Feber, Elżbieta Jurkiewicz, and Lidia Ziolkowska. Left and right ventricular morphology, function and myocardial deformation in children with left ventricular non-compaction cardiomyopathy: A case-control cardiovascular magnetic resonance study. *Journal of Clinical Medicine*, 11(4):1104, February 2022.

[68] Nada Yousif, David Niederseer, Angela Davies, Mohamad El Issa, Bilal Sidia, Hafsa A. Noor, Hamza Amin, Lorenz Räber, Baris Gencer, Robert Klingenberg, Stephan Windecker, François Mach, Christian M. Matter, David Nanchen, Thomas F. Lüscher, and Salim Obeid. Impact of malignancy on clinical outcomes in patients with acute coronary syndromes. *International Journal of Cardiology*, 328:8–13, April 2021. Epub 2020 Dec 13.

[69] Michael A Matter, Francesco Paneni, Peter Libby, Stefan Frantz, Barbara E Stähli, Christian Templin, Alessandro Mengozzi, Yu-Jen Wang, Thomas M Kündig, Lorenz Räber, et al. Inflammation in acute myocardial infarction: the good, the bad and the ugly. *European heart journal*, 45(2):89–103, 2024.

[70] Stephan Winnik, Johan Auwerx, David A Sinclair, and Christian M Matter. Protective effects of sirtuins in cardiovascular diseases: from bench to bedside. *European heart journal*, 36(48):3404–3412, 2015.

Signature

By submitting this proposal, we confirm that the scientific lead has understood the requirements for participation according to the call document, and commit to the necessary levels of lab and institutional co-funding if the proposal is successful.

Place and Date: 8.9.2025
Name and signature of the scientific lead:

Prof. Dr. Julia Vogt, ETH Zurich

A Preliminary Activities

Task	GPU. Spec.	Model Size	Single Run GPU hours	Number of Runs	Number of Datapoints	Total GPU hours
VLM Finetuning	H200	8B	1024	30 - 40	∼1,150 k	50k
VLM Evaluation	H200	8B	50	100	10k	5k
Total						55k

Table 1: Compute time utilized for the RadVLM experiments in terms of GPU hours as part of the Swiss AI health vertical project.

Task	GPU. Spec.	Model Size	Single Run GPU hours	Number of Runs	Number of Studies	Total GPU hours
CXR Pretraining (ViT-B)	H200	135 M	140	∼500	101 k	70k
CXR Pretraining (ViT-L)	H200	370 M	768	25	101 k	19.7k
CXR Finetuning & Evaluation (ViT-L)	H200	340 M	128	50	101 k	6.4k
Total						96.1k

Table 2: Compute time utilized for CXR pretraining experiments in terms of GPU hours for the Swiss AI health vertical project.

Task	GPU. Spec.	Model Size	Single Run GPU hours	Number of Runs	Number of Samples	Total GPU hours
US Pretraining (ViT-B)	H200	135 M	336	30	525 k	10k
ECG Pretraining (ViT-B)	H200	135 M	192	70	345k	13.4k
Total						23.4k

Table 3: Compute time utilized for US and ECG pretraining experiments in terms of GPU hours for the Swiss AI small grant ”Unifying Signals and Structure: A Multimodal Cardiopulmonary Foundation Model”

B Data

Table 4: Overview of CXR datasets

Dataset	**# Images**	**# Studies**	**# Patients**	**# Reports**
MIMIC-IV-CXR	377 000	227 000	60 000	227 000
CheXpert Plus	224 000	188 000	65 000	188 000
ChestX-ray14	112 000	112 000	30 000	-
PadChest	160 000	109 000	67 000	109 000
RexGradient	255 000	160 000	109 000	160 000
PediCXR	9 125	9 125	9 125	-
Total	**1 137 125**	**805 125**	**340 125**	**684 000**

Table 5: Overview of ultrasound datasets

Dataset	# Samples	# Patients	Data type
MIMIC-IV-Echo	525 000	4 579	Video
RadImageNet	390 000	N/A	Image
EchoNet	10 030	10 030	Video
US-4 Dataset	1 051	N/A	Video
CAMUS	500	500	Image
Detailed Dataset of Breast Ultrasound scans	266	256	Image
Breast Ultrasound Images Database	232	N/A	Image
Breast Ultrasound Videos	188	N/A	Video
Open Access Series of Breast Ultrasound	100	78	Image
CLUST	64	N/A	Video
EchoNet-Ped	4 476	1 958	Video
Total	**931 907**		

Table 6: Overview of ECG datasets

Dataset	# Samples	# Patients
MIMIC-IV-ECG	800 000	160 000
CODE-15	345 779	233 770
SPD	25 770	24 666
PTB-XL	21 779	18 869
SaMi-Trop	1 631	1 959
European ST-T	90	79
MIT-BIH Arrhythmia	86	80
Physionet Long-Term ST	48	47
Fantasia	40	40
BUTQDB	18	15
Total	**1 195 241**	**439 525**

Table 7: Overview of ICU,EHR, and medical notes datasets

Dataset	# Samples	# Patients
MIMIC-IV-ICU	94 458	65 366
MIMIC-IV-hosp	546 028	223 452
MIMIC-III	53 423	38 597
NWICU	28 150	25 923
AmsterdamUMCdb	23 106	20 109
HiRID	33 000	Not Disclosed
SICdb	27 386	21 583
MIMIC-IV-Note (discharge summaries)	331 794	145 915
MIMIC-IV-Note (reports)	2 321 355	237 427
Pedicatric Intensive Care (PIC)	13 449	12 881
Total	**3 472 149**	**791 253**