PROJECT SUMMARY:

I. Title

Dynamic AI foundation models for interpreting medical images through self-supervision, multimodality, clinical-context prompting, and adaptable explainability.

II. Theoretical-Conceptual Framework

We are witnessing rapid AI progress, particularly in natural language processing and computer vision [1-5]. These breakthroughs are revolutionizing various industries and hold immense potential for the healthcare sector. Healthcare is a social good whose progress directly translates to increasing patient quality of life. The use of Artificial Intelligence (AI) in medicine is not new - for decades, researchers have been working on accelerating diagnostics and decision-making to benefit patients. However, with extensive data repositories and advancements in computational capabilities and deep neural architectures, AI models have shown much higher capabilities than we ever thought possible a decade ago. This is leading us to a transformative moment in healthcare, combining physicians' cognitive abilities with AI models' power. Many experts see this combination as a winning team for advancing healthcare. However, despite these advancements, these technologies have limited in-use applications in the clinical workflow, specially in our region, where most healthcare institutions continue to use outdated protocols [10]. Most of the solutions are designed as isolated components with poor level of generalization which do not integrate to the multimodal source of information of patients. In addition, the solutions lacks of an explainability engine capable of adapting to different cases.

Therefore, **our main goal** is to propose a general AI-based medical foundation model (FM) that deals with diverse image modalities (Computerized Tomography – CT, Positron Emission Tomography – PET-CT, Whole Slide Image - WSI) and is trained with multimodal data beyond images. Our proposals should be characterized by easy adaptation to different problems, self-supervision, multimodality and explainability. In terms of the application, these models should interact with the specialist dynamically through efficient clinical-context prompting strategies. Our proposals should also interact with the electronic health records (EHR) of patients to have a holistic understanding of each situation.

III. Research Questions and/or Hypotheses, Problem Statement of the Research Proposal

AI is shaping a new era of truly personalized patient care. However, it is crucial to recognize the need for ongoing research and rigorous validation to ensure the safe and effective implementation of AI in healthcare. Multidisciplinary collaborations involving specialists from AI, radiology, pathology, and clinical practice are essential for refining AI-driven models and technologies. Additionally, we still need to address numerous challenges to fully realize the impact of AI in the medical field, such as multimodality explainability, generalization, adaptability, and appropriate interaction mechanisms.

To address the mentioned challenges, we propose a dynamic AI medical FM for interpreting medical having the patient as the center of our proposal. Here, the term dynamic means that the models can be easily adapted to different radiological/histopathological context. Features like clinical-context prompting and adaptable explainability should be aware of patient benefits. Thus, to achieve our goal, we need to answer the following questions: What is the most efficient method to leverage diverse health datasets to produce a dynamic medical FM? Would training a medical FM under a multimodal regimen allow it to improve its performance in medical image interpretation? Is it possible to align data from diverse sources to train a FM in the medical imaging context? Is it possible to provide a general medical model to deal with histopathology and radiology imaging together, working as one? What are the best prompting strategies for physicians to interact with a medical FM for interpreting medical imaging? How should we evaluate the generalization property of AI FMs for interpreting medical images? What should be the basis for proposing a general explainability model for medical imaging? Is it possible to have a general explainability model for the medical context interacting with the general predictor model? How do we build an inference graph from a general AI FM for interpreting medical images?

In this sense, we claim the following hypothesis: self-supervision, multimodality, clinical-context prompting, and adaptable medical explainability model together allow us to propose a general AI foundation model for interpreting medical images characterized by a high level of generalization.

IV. General Objective

To propose a general dynamic AI foundation model for interpreting medical images with focus on radiology and histopathology, leveraging a diversity of imaging modalities trained under multimodality and self-supervision, including an adaptable explainability module and clinical-context prompting strategies. In this project we limit our study to fine and coarse grain segmentation and classification of tumor mutation, as well as cancer grading.

V. Specific Objectives

(SO1) To build **multimodal datasets** for training FM under self-supervision addressing the problem of non-aligned data to evaluate SOTA and proposed models.
(SO2) To design a **general architecture** of the multimodal AI FM and evaluate the performance on a set of defined target tasks.
(SO3) To design clinical-context **prompting** strategies for dynamic interaction optimized for the medical context.
(SO4) To design an adaptable medical-imaging **explainability** model that can be integrated into the proposed FMs.
(SO5) To **assess** the proposed models **in real scenarios**, particularly, in radiology and histopathology.

VI. Methodology

SO1: To build multimodal datasets for training foundation models under self-supervision addressing the problem of non-aligned data

We will leverage the diverse publicly available datasets proposed together with IA models in the medicine field. For instance, in the case of CTs we have already collected 24.000 thorax CT studies representing up to 1.6 millions of images. We also leverage our cooperation with local health institutions to increase our multimodal datasets. Indeed, we will access to a large amount of CTs, PET-CTs, WSIs studies, but most of them are isolated studies which are not connected each other. Isolated data is not enough, we need a mechanism to build a multimodal repository with data coming from different sources to be aligned connecting different modalities each other according to a set of heuristics. To this end, we leverage the metadata available for each study, which include information of pathologies that is valuable to connect diverse studies.

Our proposed dataset will allow us to have a baseline for future comparisons for the AI-based medical imaging community as a reference point. With respect to generalization, we are interested in assessing this property under the following contexts: new-task and cross-setting generalization.

Activities: (1.1) Propose a taxonomy of the different existing datasets, (1.2) evaluate fairness/bias in the existing datasets. Here we will work in a demographic study of the available data, (1.3) design alignment mechanisms to join different data modalities. We have a diverse data source; it is important to find mechanisms to join these sources, (1.4) define protocols to use the datasets and make them public, (1.5) define evaluation criteria and involved metrics with respect to new-task and cross-setting generalization, (1.6) define evaluation datasets taking into account fairness and diversity among medical image modalities, and (1.7) run benchmarks according to the model typology with focus in generalization cross-tasks and cross-settings. These activities will take place during the year 1.

SO2: To design a general architecture of the multimodal AI foundation model and evaluate the performance on a set of defined target tasks.

This is the core of our project because the result will be the main architecture of our proposals. We start with an initial proposal which brings many challenges related to the design of the following components:

- **Encoder:** this is the responsible of translation of the input into semantic representations (encodings). Different from traditional *visual encoders* like Dinov2, ours must allow multimodality. Thus, we need to extends ideas similar to CLIP on multiple data sources to produce encoders for each kind on input modality.
- **Decoder:** this takes the latent vectors to solve target tasks. The objective is to build a general decoder that can be dynamically adapted through a limited different tasks by prompting mechanisms. The scope and taxonomy of the tasks that can be modeled by one decoder should be answered by the activities involved in this task. However, we propose the following three groups of tasks, where each can be addressed by one decoder: (1) cancer scoring on breast and prostate WSI, (2) fine and coarse grain liver segmentation, and (3) classification of lung tumor mutation (EGFR/KRAS).
- **Prompting:** the goal is to encode different prompting mechanisms to guide the decoder in target tasks. Considering the challenge on defining appropriate clinical-context prompts, we will go in depth about this problem in the following objectives.
- **Explainability:** this is a critical component which will provide a task-dependent explainability of the model, in the form on natural language using LLMs and considering the Ordinary Language Philosophy. This component is also described in another objective later.

Activities: (2.1) Design the general architecture for our medical-imaging FMs, (2.2) implement and train the encoder under self-supervision using multimodal data, (2.3) define the method to train the decoder using clinical-context prompting, (2.4) evaluate our proposals with respect to metrics described in SO1, (2.5) characterize the tasks that can be solved by the same decoder. These activities will take place during the year 2.

SO3: To design clinical-context prompting strategies for dynamic interaction optimized for the medical context

Prompting is a versatile way to interact with a FM to modulate its inference process. Our objective is to provide a general model that can be easily adapted through a diverse set of prompting mechanisms. Unlike FM for LLM, where the prompts are provided for a diversity of persons, in our case, the users will be physicians or health professionals. Therefore, we must design proper prompting modalities matching the healthcare workflow. In addition, we will explore new prompt modalities beyond simple text. In this vein, we will evaluate sketch-based and stroke-based prompts. To this end, it is critical an interdisciplinary work between the AI researchers and the healthcare personnel.

Activities: (3.1) Explore the nature of prompting in the clinical context, (3.2) design an evaluation methodology for prompting strategies in the medical domain, (3.3) design text-based clinical-context prompting and evaluate the value to medical domain, (3.4) design sketch- and image-based clinical-context prompting and evaluate the value to medical domain, and (3.5 evaluate the clinical-context prompting strategies integrated to the proposed FM. These activities will take place during the third year.

SO4: To design an adaptable medical-imaging explainability model that can be integrated into the proposed foundation models.

Explainability is a research topic with very few results in the medical context. So far, most proposals are based on heat and saliency maps, which, contradictorily, do not provide much explainability to the health personnel. Even though, explainability is an unsolved problem in medical context, our proposal adds a more challenging feature. We are interested in a general dynamic explainability model, able to be adapted to different problems. Our initial idea is to build an inference graph representing the model's path activation. Each graph node should group a set of neurons activated by similar features. Our explainability model must learn a relation between the nodes of the inference graph and the input. All the nodes are called evidence, and the edges are the dependencies between evidences. To measure the quality of explainability we will use an objective, deterministic and model/agnostic algorithm, known as DoX. Thus, we need to define, as prompts, a set of aspects to asses, a set of archetype queries and a task.

Activities: (4.1) explore explainability strategies in the medical field, (4.2) evaluate state-of-the art explainability strategies regarding on the opinion of experts, (4.3) design and implement the dynamic explainability model, (4.4) evaluate the dynamic explainability model according to DoX, and (4.5) integrate the explainability model into the FM. These activities will take place during the years 3 and 4.

SO5: To assess the proposed models in real scenarios, particularly, in radiology and histopathology.

A special characteristic of our proposal is that we are motivated by the impact that our work brings into real scenarios. To this end, beyond public datasets we need to evaluate the real impact in other datasets, particularly obtained from LATAM. Therefore, to this end we leverage the cooperation of our health team to collect diverse datasets to evaluate our proposal under the same metrics as defined in SO1.

(5.1) Collect WSI of breast cancer biopsies for HER2 scoring, (5.2) collect WSI of prostate cancer biopsies for Gleason scoring, (5.3) collect abdominal CT for hepatic vessel segmentation, (5.4) collect thorax CT, PET-CT, WSI of lung cancer biopsies for classification of lung tumor mutation, and (5.5) evaluate our proposed FM on the collected data. These activities will take place in year 4.

VII. Justification of the following points:

i. Description of the high novelty and disruptive nature of the proposal.

Foundation Models (FM) represent the most recent advances in deep learning, marking the current moment of modern machine learning. In the context of medical imaging, some proposals have attempted to represent general models like UNI[6], MedCLIP[7], BiomedCLIP[8], or MI-Zero[9]. However, these proposed models are still far from real medical FM. They only include bimodality combining text and images, but they do not incorporate general multimodality. In addition, they do not include explainability or dynamic interaction. Another critical drawback is the strong disconnection between the AI specialists and medical staff, which promotes the proliferation of AI models with minimal or no impact.

Therefore, the novelty of our proposal is the design, implementation and release of the first general AI-based medical foundation model characterized by multimodality, self-supervision, dynamic interaction by clinical-context prompting and adaptable explainability. Each of the characteristics of our model brings novelty per se, as described below:

- **Multimodality**: current proposals do not include multimodality beyond combining text with images. Our proposal leverages different source of patient’s data like medical images (CT, PET-CT, WSI), reports and EHR to have a holistic and accurate understanding of the underlying situation.
- **Self-supervision**: most of the models are based on supervised learning requiring abundant “gold-

standard" labeled data. In contrast, our model is based on directly exploiting existing data through self-supervision mechanisms. To this end, we will propose alignment methods to relate the diversity of available data.

- **Clinical-context prompting**: a prompting mechanism allows us to interact with an AI model without technical knowledge. The use of prompting is popular in LLM, where commonly text or images are used as prompts. However, there is no studies about the appropriate prompting strategies in the clinical context. Therefore, part of our work will be focused on studying and proposing efficient prompting mechanisms to facilitate interaction with a medical general AI model.
- **Adaptable-explainability**: this is the most disruptive feature of our model. Current FM are based on predictions without including any explanation about the inference. In the medical context, explainability is critical and depends on many factors like the underlying pathology or the patient's EHR in general. In addition, a predictor model is useless without a solid explanation. Therefore, this work proposes an adaptable explanation model to be integrated into the medical foundation model. The explainability model should modulate its behavior depending on the used prompts.

Considering our proposal's disruption and innovative level, there is a higher level of uncertainty in diverse aspects:

- Even though we could access to a large amount of medical data, most of them are unimodal, thus alignment methods to appropriately connect then are required. Therefore, there is an evident level of uncertainty about the effectiveness of our alignment strategies to hold coherence and representativity in the underlying problem.
- One of the key uncertainties lies in our models' ability to handle diverse image modalities, such as CT, WSI, PET-CT images. There is a potential inflection point where the addition of more modalities could lead to a collapse in the model's performance, posing a significant risk to our project's success.
- There is uncertainty about the level of generalization we can reach with our FM. It is utopic thinking to propose a full general model for every possible task. We aim to propose the most general foundation model possible. To this end, we need to characterize the group of tasks for which our general model can run successfully.
- Uncertainty also surrounds the level of interaction we can achieve with our prompting strategies. While we believe that text and sketches can effectively engage physicians in their daily activities, it's crucial to acknowledge that these initial proposals may not be foolproof. Therefore, we propose a collaborative approach, creating an ecosystem that fosters close cooperation between AI researchers and physicians to develop more efficient prompting mechanisms.
- Uncertainty exists about how we can align diverse image modalities to allow them to interact and power up high semantic representation. Our first idea is to exploit metadata about pathologies to align diverse modalities. However, if this approach does not work, we will need to explore other alignment mechanisms. For instance, we can use existing predictors to match data with based on similar inference.
- There is high uncertainty about the level of adaptability our explainability model can achieve. We propose an inference-graph explainability model, but we must determine whether our model can handle diverse medical imaging tasks.

ii. Description of the transformative potential in the research field and/or the potential benefits of its results or research process.

Our proposal introduces a groundbreaking approach to medical imaging. We will be pioneer developing a general AI foundation model that interprets medical images using multimodal data, thereby minimizing bias and enhancing sensitivity. These models will be adaptable to various contexts and settings, fostering interaction through meticulously designed clinical-context prompting strategies. Moreover, we will propose an innovative explainability module to be seamlessly integrated into the foundation models. The adaptability of the explanation model will be contingent upon the prompting and the predictor's activations, marking a significant advancement in AI-driven medical image interpretation.

By integrating diverse patient data sources such as medical imaging (CT, PET-CT, WSI) and EHRs into a general AI foundation model, our proposal will revolutionize healthcare. This integration will pave the way for personalized treatment plans, precise disease risk prediction, and enhanced monitoring of treatment responses. The proposal will also elevate clinical decision-making by amalgamating patient-specific data with the latest scientific discoveries to generate pertinent recommendations. Importantly, all these advancements will be patient-centric and adhere to ethical considerations such as data privacy, security, and algorithmic bias. The multidisciplinary collaborations envisaged in this proposal, involving specialists from AI, radiology, and pathology, are crucial for refining the AI-driven general foundation model for interpreting medical images with a direct and positive impact on the patient's journey.

Our general model holds immense potential to deliver significant value in the clinical context. It will drive down costs by enhancing efficiency through a more streamlined workflow and shorter image reading time. Simultaneously, it will enhance patient health through early detection, reduced radiation and contrast agent dosis, improved diagnostic accuracy, and personalized diagnostics. These benefits underscore the feasibility and value of our proposal, making it a compelling choice for healthcare providers and stakeholders.

VIII. References

[1] Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763, 2021.

[2] Jean-Bastien G., et al.: Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. Annual Conference on Neural Information Processing Systems (NeurIPS), 2020.

[3] Caron, M., et al.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9630–9640, 2021.

[4] Brown T., et al.: Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[5] Touvron, H., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[6] Chen, R., et al.: Towards a general-purpose foundation model for computational pathology. Nat Med 30, 850–862, 2024.

[7] Wang, Z., et al.: MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. EMNLP, 3876-3887, 2022.

[8] Zhang, S., et al.: BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv:2303.00915, 2024.

[9] Lu, M., et al.: Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images. CVPR,19764-19775, 2023.

[10] Han, R., et al.: Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. The Lancet Digital Health, https://doi.org/10.1016/S2589-7500(24)00047-5, 2024.