PROPOSAL ABSTRACT

Name of Principal Investigator:	José Manuel Saavedra Rondo
Proposal Title:	Leveraging Visual Latent Space and Large Language Models to Facilitate Interpretability and Explainability of Vision Models

Although we have seen a tremendous impact of deep learning models in diverse image-based applications, a critical problem still hinders the mass adoption of image-based solutions in areas related to human risks, such as medicine, autonomous driving, or law-related applications. Current models are mainly used as **black boxes** (opaque models), lacking reasoning behind the inferred outcomes. Furthermore, the increasing size of foundation models, although effective, makes the reasoning process harder. Thus, the necessity of models capable of reasoning about outcomes has led to the emergence of what we call **Explainable Artificial Intelligence** (XAI), the core topic of this project.

An **explainable model** is a computational model aiming to provide specific details or reasoning regarding its behavior to produce certain outcomes. Interpretability and explainability are frequently used interchangeably; however, they refer to distinct aspects of an XAI model. **Interpretability** is a property of an XAI that focuses on the inherent clarity of a model's structure and how it makes decisions (e.g., a decision tree), while **explainability** aims to provide justifications or reasoning for specific predictions. Explainability is complex in that it depends on the specific application task, which makes generalization very challenging. For instance, providing a medical diagnosis based on the occurrence of specific symptoms represents an explainable prediction. Current methods based on neural networks are far from being explainable as they are solely trained to produce the correct answer without taking in account the inner behavior of the model.

The **encoder-decoder architecture** is the de facto standard in modern AI, particularly in the image domain. The encoder is responsible for producing highly semantic representations (embeddings) and may cover up to 80% of the total parameters of a complete model. The decoder seeks to transform the visual embeddings into final predictions. Moreover, current **foundational models** for images are focused on improving visual representation, such as DINO-V2, iBOT, or CLIP, which are typically known as visual encoders.

Considering the importance of visual encoders in the process of making predictions, SOTA XAI models for images are based on fusing visual embedding with LLMs such as DDCoT or SPANet. LLMs are critical for facilitating rationales about an image, which bypasses the limitation of having a massive amount of labeled datasets connecting images with rationales.

Although some works addressing explainability for images have appeared, there are still strong limitations that need to be addressed. For instance, current models do not exploit multiscale representations provided by visual encoders; they are mostly based on supervised learning which limits the generalization of the models; they are based on a quadratic-complexity attention which limits the context size and scalability; and they are based on producing textual rationales where visual explanations could fit better for specific domains.

Therefore, this work aims to address the above limitations, proposing a set of XAI models based on the hierarchical analysis of the visual feature space and the knowledge of LLMs to provide rationales. Our proposal will incorporate the best aspects of the different inherently explainable mechanisms to push the frontiers in the underlying area. Additionally, we propose moving toward visual explanations by leveraging advancements in diffusion-based generative models.

To achieve the general objective, we propose the following specific objectives: **OE1)** to suggest a mechanism to select multiple representation scales from visual encoders. To this end, we leverage the prototype-based and concept-based models. **OE2)** to define a self-supervision mechanism to generate rationales from SOTA LLMs that can be aligned with multiscale visual representations. To this end, we will extend the DDCoT proposal, including concepts and prototypes. **OE3)** to propose a hybrid XAI model merging concept-based and prototype-based strategies with chain-of-thought mechanisms under self-supervision and multiscale representation. **OE4)** to propose a new XAI modality based on visual explanations leveraging conditional diffusion models. **OE5)** to evaluate our proposal in real environments, particularly in areas related to human risks from simple to more complex tasks.

We present a 4-year plan, focusing on the first three objectives during the first two years of the project. Thus, at the end of the second year, we will have our improved XAI model tested on small scenarios. The third year is devoted to proposing a challenging visual explanation model that leverages our experience with diffusion models. Finally, the fourth year will be focused on evaluating our proposal in different domains with the support of our collaborators. As a consequence, we hope to publish at least four papers by the end of the project and contribute to the formation of advanced human resources.

Furthermore, it is essential to highlight the main researcher's experience in the area of deep-learning models for diverse vision tasks. His publication records and awarded funds support him. Additionally, he advises on the collaboration of various

specialists where XAI models offer significant benefits.