Taming Vision-Language Models for Explainable Video Understanding for Human-Robot Interaction

Naveen Vellaturi, UNC Charlotte
Srijan Das, College of Computing and Informatics

Introduction

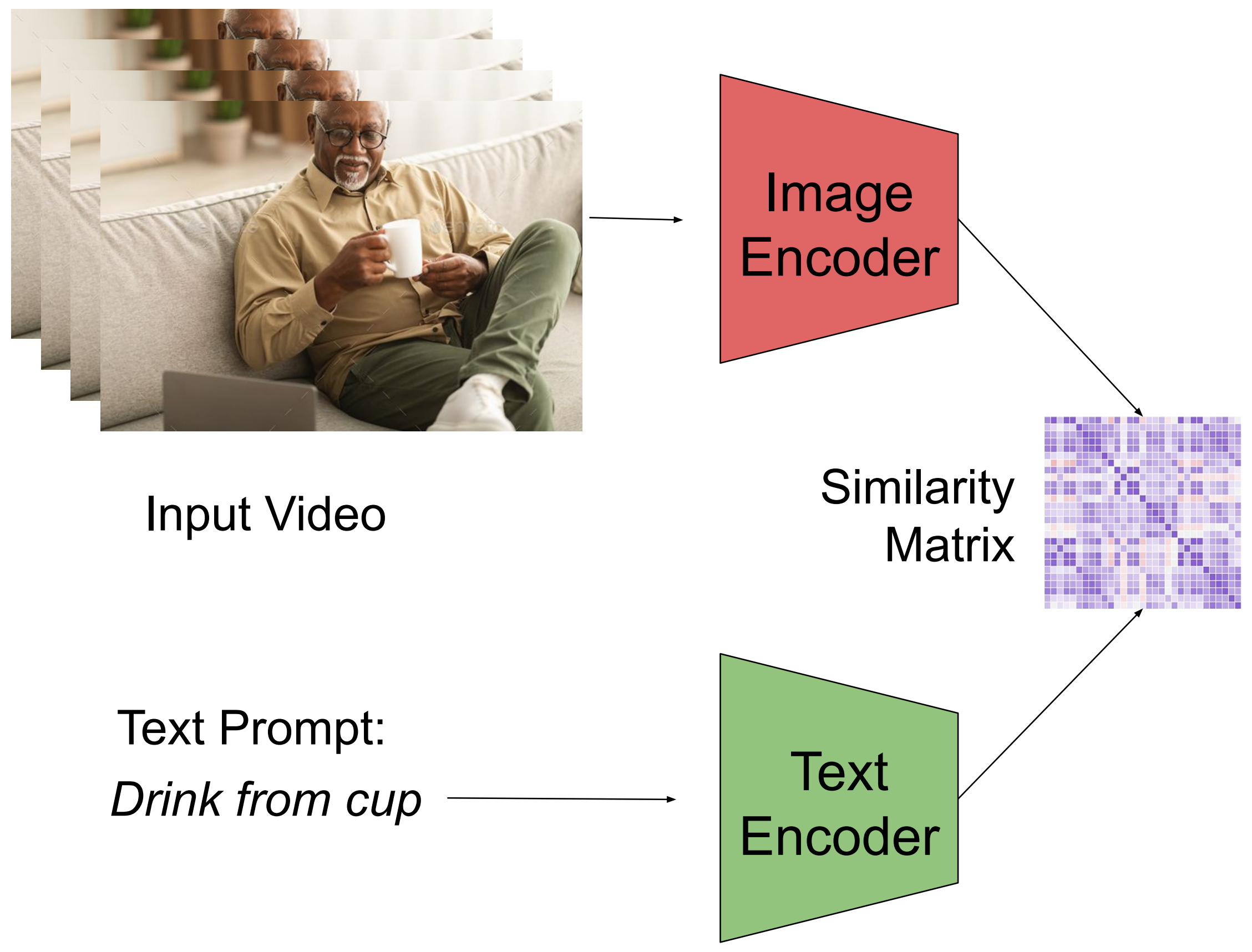
- An AI-powered industrial robot's error led to a fatal accident when it misclassified a person as an item. This highlights the importance of AI trustability when it comes to safety.
- A **lack of transparency** in these models **raises safety concerns** despite accurate action predictions.
- **Understanding a model's rationale is imperative** for safety and reliability.
- Our goal is to improve a model's explainability by having a model tell us what **attributes** of a video it uses to make its classification.

Objectives

1. **Develop an interpretable video model for AI applications.**
2. **Prompt an LLM to generate descriptive *attributes*.**
3. **Develop a Vision-Language model to extract attributes per frame**
4. **Develop a linear function to learn the mapping from the attributes to the actions.**

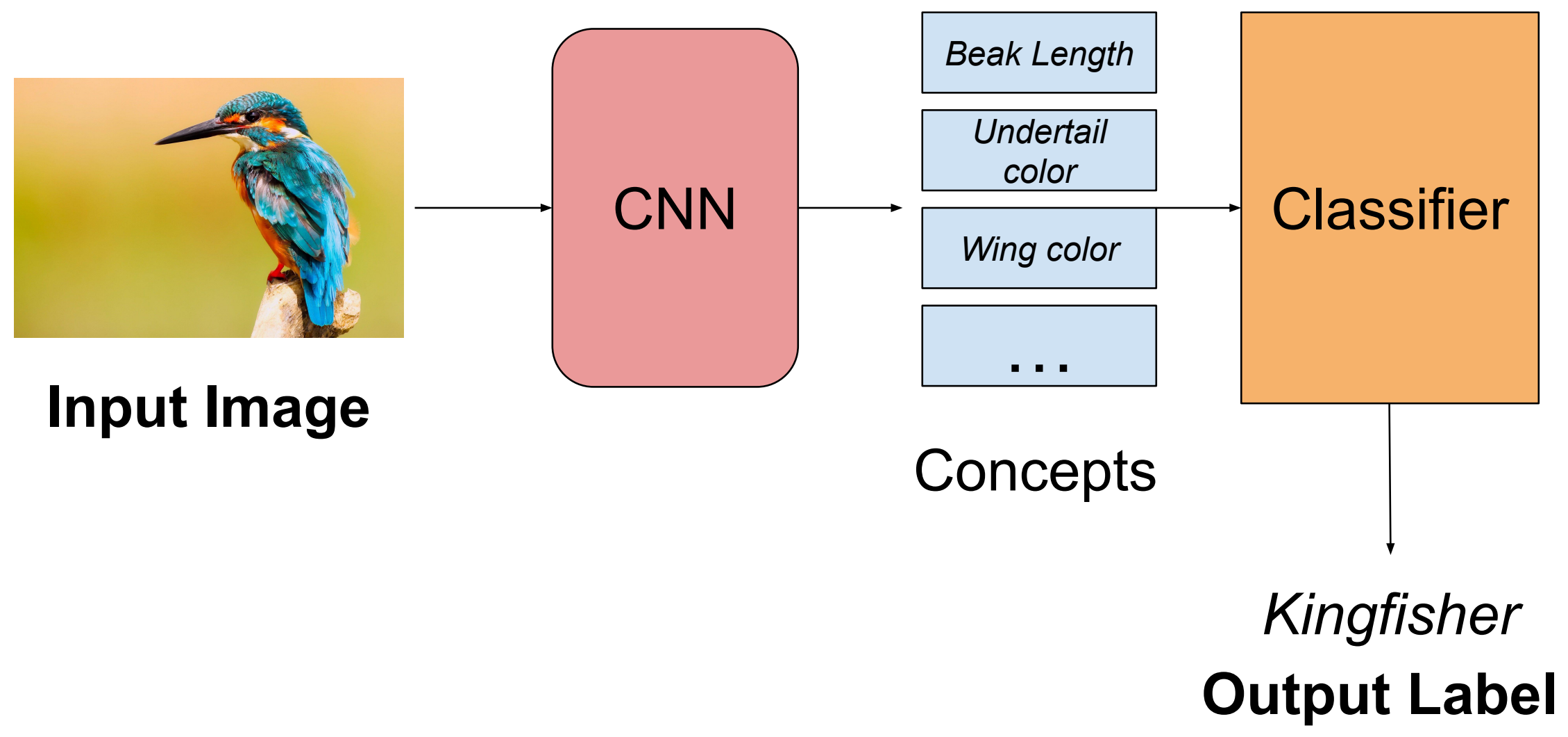
Prior Work

Vision Language Models



- In Vision Language Models, the image encoder produces frame-level representations which are aggregated, and the text encoder is responsible for the text embedding of the corresponding action label.
- The model aims to maximize the similarity of the respective representations.
- Image Encoder can have a Transformer or CNN backbone.

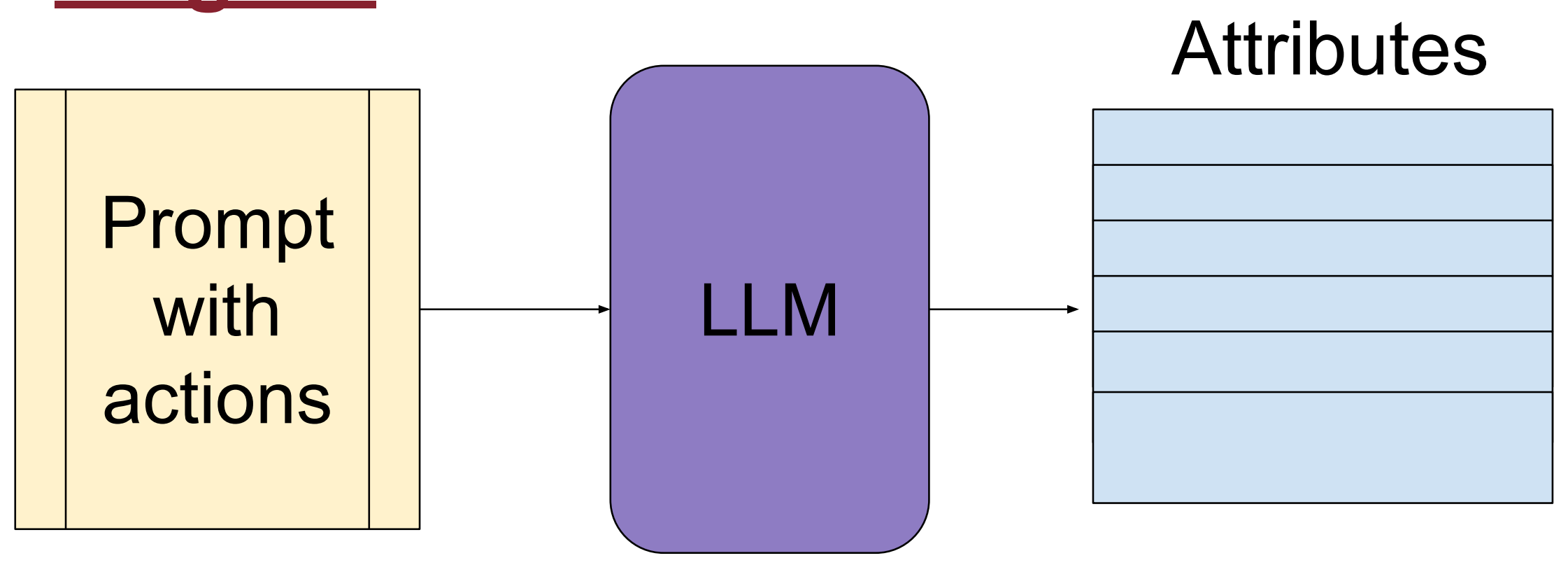
Explainable Models



- Concept Bottleneck models first use an image encoder to predict a set of concepts.
- These concepts are then used as an input to a classifier for image classification.
- These intermediate concepts help users to easily understand the model and to interact with it.

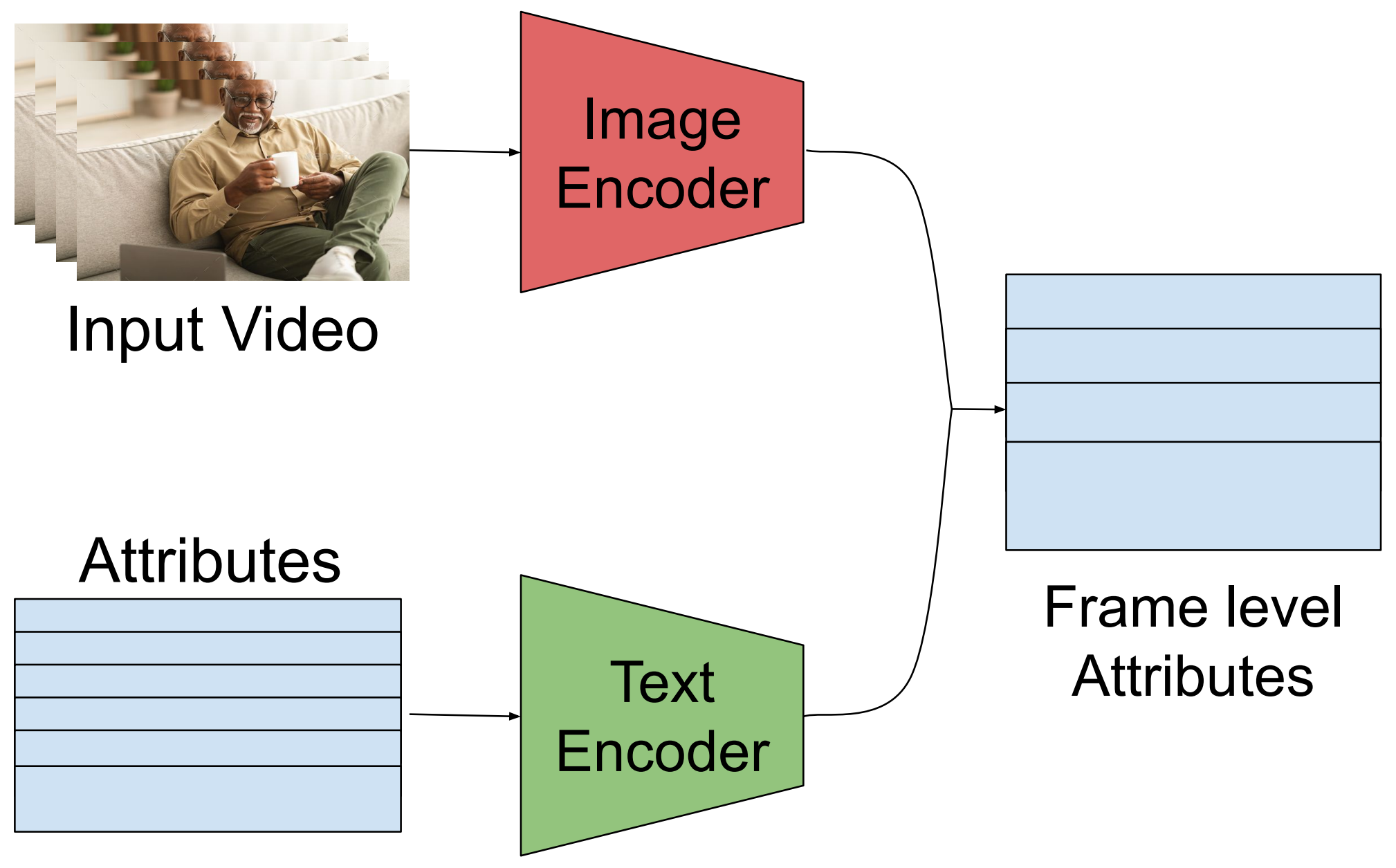
Our Proposal

Stage 1:



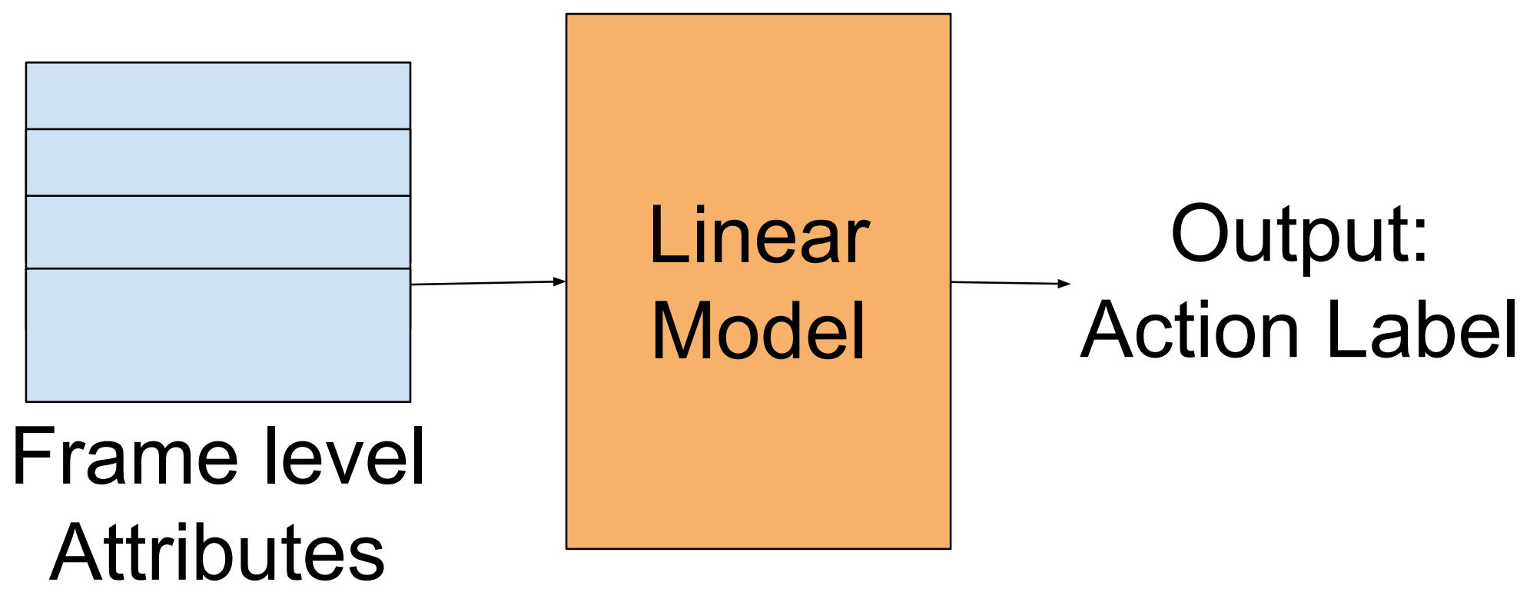
- We prompt an LLM like GPT 4 to produce attributes for each action.
- **Attributes** are the parts of an action - what a person or a model would use to to make a classification.

Stage 2:



- Previously generated attributes and video frames are used to generate a similarity matrix.
- Matching attributes are outputted for each frame.

Stage 3:



- Attributes are put through a linear model with **no nonlinearity** (maintains explainability) to produce the action label.

Conclusions

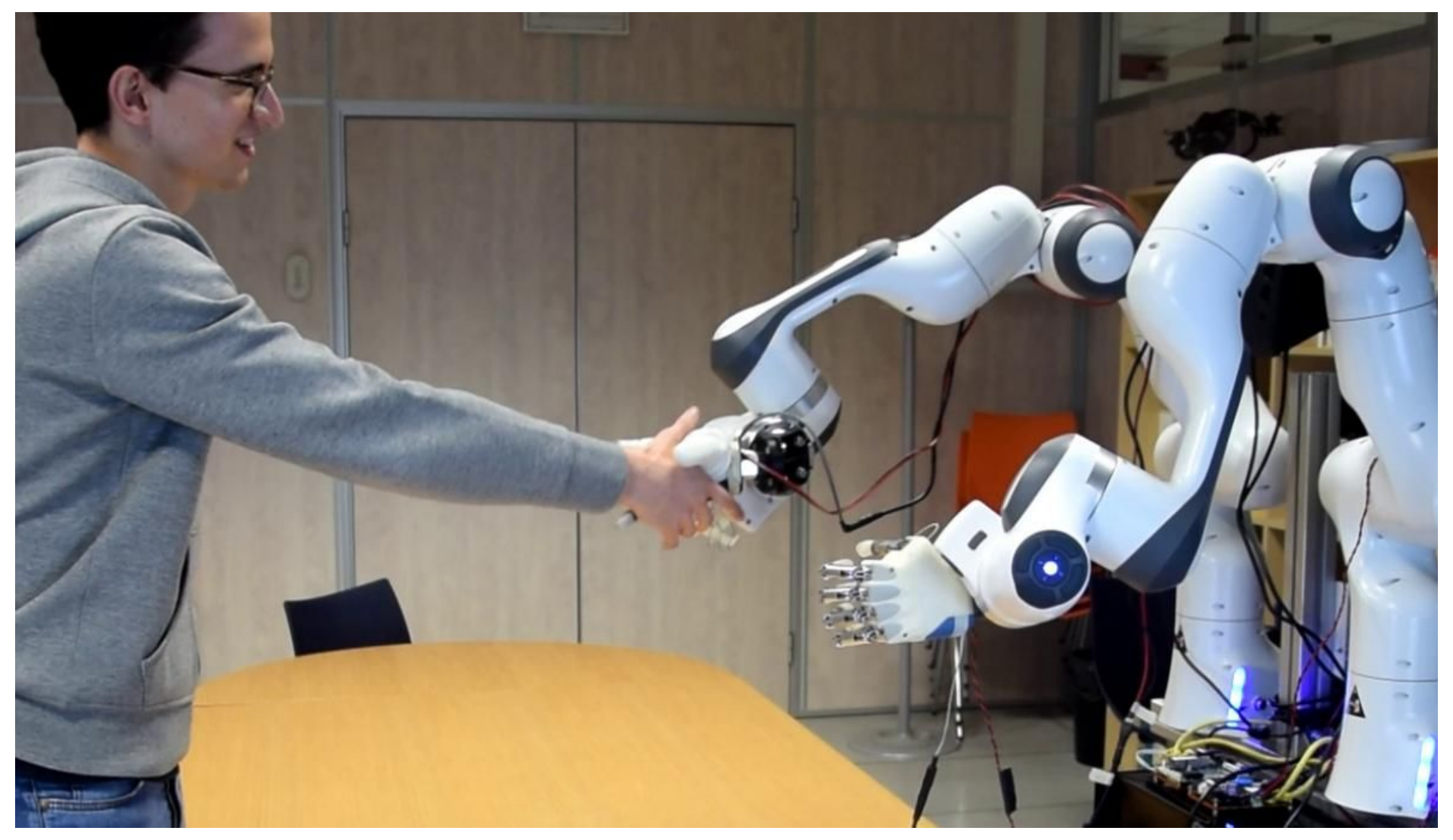
- Our proposal is currently in progress.
- We expect the accuracy to be less than state of the art Video Language Models.
- However, our model has the benefit of explainability.

Explainability… Why?

- It results in a **self-interpretable video model** with a linear mapping from per-frame concepts to its predictions.
- Users can **understand the model's reasoning** for it's prediction, making the model more trustworthy.
- User can **correct any errors** in the attribute prediction to get a **better final action prediction**.
- Users can **prevent errors** in high stakes environments.

Applications

- Monitoring the elderly in smart homes.
- Patient care in healthcare settings.
- Safe Human-Robot interactions.



References

Dai, Rui, et al. "AAN: Attributes-Aware Network for Temporal Action Detection." arXiv preprint arXiv:2309.00696 (2023).
Rasheed, Hanoona, et al. "Fine-tuned clip models are efficient video learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.