Summary. My main research interests focus on designing **temporal probabilistic deep learning models**, which describe uncertainty in trajectory prediction and generation, motivated by applications in computer vision as well as other multi-modal tasks. My portfolio consists of more than 10 papers and patents, with some of the work published in top-tier conferences like CVPR and UAI. During my PhD, I interned in several research labs, such as Amazon Alexa AI, Microsoft Research (MSR), and NEC Labs America, where each project led to a first author paper. Next, I will outline my ongoing/future research interests and discuss how my past research experience can help in moving forward.

1 Current research

I worked on several different aspects of machine and deep learning, but one area that particularly interests me is the temporal nature of data. I have explored different types of temporal processes, such as sensors information (e.g., audio), coordinates representing physical positions of the agent, 2D images, 3D tensors corresponding to brain-imaging scans, and so on. For images, I had to leverage not only the temporal but also the **spatial-temporal information**. I worked on a variety of tasks, such as interpolation (e.g., frames generation between observed scenes), extrapolation (e.g., reconstructing a brain image into the future), trajectory generation (e.g., synthesis of new behaviour of an agent), and classification (e.g., whether a sequence of actions is a ransomware attack).

In contrast to other works that consider temporal processes as discrete, my approach uses a continuous perspective. This view allows for a more accurate representation of the underlying processes and enables more effective analysis and interpretation of the data. Specifically, I was inspired by **Neural ODE** (NODE), Duvenaud's view that in data space, we observe discrete outputs from an underlying continuous process, which can be represented using a Differential Equation (DE). Once we characterize a DE, we can produce an output at any time step, which makes them well suited for interpolation and extrapolation tasks. For example, to make a smooth interpolation of non-existing frames, I introduced a generative model, **Warping NODE**, which models the vector field of changes (warps) between frames via NODE.

A related (but still distinct) aspect of my work is the use of **probabilistic modeling**, which ranges from Bayesian Neural Networks, to other means of incorporating uncertainty in a model. Departing from the existing Neural ODE literature, which assumes that a single trajectory is initiated from an initial condition, I have suggested in several papers different architectures (**Mixed Effects NODE** and **Functional NODE**), where multiple trajectories can be initiated from the same initial point. This idea was originally motivated by the observation that the brains of two twins may develop differently over time, despite having the same genes and starting at the same point in their development. Such probabilistic approaches allow us to model uncertainties around the generated trajectories and provides a mechanism for sampling and generating multiple possible trajectories for an agent. This can help to improve the accuracy and robustness of our predictions and enable us to better understanding of the agent over time.

During my internship at Amazon Alexa AI, I conducted research on **Vision Language Models** (specifically VQA), which sparked my interest in the temporal aspects of multimodal datasets, discussed further in the next section.

2 Future research: multimodality angle.

The multimodal nature of temporal datasets poses many interesting questions, where I believe my expertise will offer novel lines of attack. I will provide more details next.

Fusion. The highly non-linear relationships among low-level features across different modalities make it challenging to determine the appropriate diffusion moment and representation for these temporal processes. For example, to accurately capture the characteristics of each modality, it may be necessary to use a short-term attention model for some modalities and a continuous Neural ODE for others, or only use a part of the temporal information. *Another example* involves controlling an agent's trajectory with a textual non temporal prompt – what would be the ideal mechanisms to condition on it? What role do human-in-the-loop paradigms play in this setting and which statistical criteria should be leveraged?

Uncertainty/Robustness. How do different modalities, such as video, audio, and text, affect the uncertainty of predictions in both interpolation and extrapolation settings? Can we disregard some modalities if they do not contribute to the prediction, or will removing modalities lead to unstable model performance? In particular, I am interested in the role of uncertainty in video understanding, and how captioning can systematically seek human guidance as well as other information such as audio/comments.

Generation of data from multiple modalities. There are already models that generate video based on a textual prompt, but it would be interesting to see models that can generate data from multiple modalities, such as video and audio, simultaneously. From a computational perspective, it is interesting to explore whether the concept of KL divergence, which is used in VAE-based generative models, still applies and if adjustments should be made based on the number of generated modalities.