**Graduate Research Experience Course**

Research Plan

Student Name: **David Lelis**  
Research Experience Advisor: **Dr. Ayan Dutta**

1. **Summary: Vision-Language-Assisted Multi-Agent Collaboration Using RL**

The goal of the project is to design a multi-agent system where agents use a vision-language model for shared situational understanding and reinforcement learning for collaborative task execution. Our research will equip agents (e.g., humans and robots) with Vision-Language Models (VLMs) to process visual inputs and textual instructions in parallel, enabling them to communicate and coordinate effectively.

We plan to use deep reinforcement learning to train agents on cooperative tasks, such as object manipulation, search-and-rescue, or surveillance in a 3D environment (e.g., an office building – UNF SoC 3rd floor). We plan to introduce a shared "language space" where agents exchange learned insights (e.g., “a red bag is found in front of Prof. Dutta’s office”) to enhance teamwork. Simulation platforms like Unity ML-Agents or Mujoco can be used to design collaborative tasks, with benchmarks such as ALFRED [1] for performance metric evaluation.

1. **Milestones**
   * **GRE I Course (Spring 2025)**

* Read and learn about Large VLMs and their application in navigation tasks (from research papers and online tutorials).
* Understand VLM models' pre-training, fine-tuning, and prompting strategies.
* Implement a vision-language model to extract shared visual and textual features for agents.
* Understand the intricacies of combining VLM and RL models. Develop and train RL policies for individual and joint agent behaviors.  
  + **GRE II Course (Fall 2025)**
* Develop a communication mechanism for agents to share insights from VLM.
* Evaluate system performance on tasks requiring coordination, such as “Find my bag that I left in the library.”
* Submit a research paper to a conference (of Dr. Dutta's choice) with the collected results.

**4. Expectations**

* (outcome) An efficient multi-agent collaboration framework in dynamic environments while finding novel insights into integrating VLM-based communication with RL-driven decision-making.
* (meetings) Meet with the advisor every week.
* (paper submission) Submit at least one paper at the end of the GRE sequence to a conference (e.g., HRI, ICRA, or IROS).

**5. Conclusion**

In this project, we aim to advance the intersection of vision-language models and reinforcement learning by developing a multi-agent system capable of collaborative problem-solving in dynamic environments. By leveraging vision-language models for shared situational understanding and RL for decision-making and coordination, the proposed framework seeks to enhance multi-agent communication and efficiency in complex tasks. The research outcomes are expected to provide valuable insights into integrating multimodal understanding with reinforcement learning, paving the way for more capable and adaptive AI systems in real-world applications such as search-and-rescue, robotic collaboration, and autonomous exploration.

**References**

1. Shridhar, M., Thomason, J., Hu, R., Matuszek, C., Sinapov, J., Blukis, V., Feng, S., & Mooney, R. (2020). "ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.