HALO: High Autonomous Low-SWaP Operations

Team Members:

- Sloan Hatter shatter2022@my.fit.edu
- Blake Gisclair bgisclair2022@my.fit.edu

Faculty Advisor: Dr. Ryan T White rwhite@fit.edu

Client: Dr. Ryan T White rwhite@fit.edu

Progress Matrix of Milestone 5

Task	Completion %	Sloan	To Do
Showcase Poster Draft	90%	100%	Add in results and data images
Obtain a CNN model trained on the same dataset as the ViT	100%	100%	None
Compare the performance of a 32- & 16-bit CNN model to the performance of the 32- & 16-bit quantized ViT model	70%	100%	Gather CNN performance metrics for comparison
Fix 8-bit ViT model	70%	100%	Prime 32-bit ViT model with Q/DQ nodes for 8-bit quantization on the Jetson
Attempt 4-bit quantization	70%	100%	Run specialized commands for 4-bit representation on the Jetson using the primed 32-bit ViT model

Discussion of Accomplished Tasks for Milestone 5:

- **Showcase Poster Draft**
  - I have drafted up the showcase poster. For the final milestone, I need to thin out the Abstract section and add to the Methods section, as they are both combined at the moment. I also need to add the results section and images of detections.
- **Obtain a CNN model trained on the same dataset as the ViT**

    - There is a CNN model that was trained by another set of researchers for vision tasks. This is the model that we are now using for comparison with our Vision Transformer model on the same kind of vision tasks.
- **Compare the performance of the 32-bit CNN model to the performance of the 16-bit quantized ViT model**
    - The new goal is to compare the fully trained 32-bit CNN model performance with our quantized 16-bit Vision Transformer model to show that the ViT model offers more accuracy without any drops, while also being more compact than the full 32-bit CNN model.
- **Fix 8-bit ViT model**
    - Successfully inserting Quantized/Dequantized (Q/DQ) nodes into the 32-bit onnx file should enable me to utilize TensorRT's INT8 command line arguments to effectively quantize the 32-bit ViT down to an 8-bit representation.
- **Attempt 4-bit quantization**
    - Similarly to above, once Q/DQ nodes are successfully inserted into the 32-bit onnx file, I should be able to use TensorRT's INT4 representation to quantize the 32-bit ViT down to a 4-bit representation.

Discussion of Contribution to Milestone 5:

- **Sloan Hatter:**

Task Matrix for Milestone 6:

Task	**Sloan**
Finalize Showcase Poster	100%
Demo Video	100%
User/Developer Manual	100%
Finalize performance metrics and results for both the quantized 16-bit ViT and the 32-bit CNN	100%
Record metrics for quantized 8-bit ViT for CNN comparison if feasible	100%
Record metrics for quantized 4-bit ViT for CNN comparison if feasible	100%

Discussion of Planned Tasks for Milestone 6:

- **Finalize Showcase Poster**
    - I will complete the showcase poster by adding finalize results and images.
- **Demo Video**
    - A demo video/simulation will be completed to show detections.
- **User/Developer Manual**
    - I will write up a user manual for how to use the Vision Transformer for deployment.

- **Finalize performance metrics and results for both the quantized 16-bit ViT and the 32-bit CNN**
    - Final performance metrics between both models will be recorded and reported as the main end goal of this project.
- **Record metrics for quantized 8-bit ViT for CNN comparison if feasible**
    - If time allows and I can achieve an 8-bit ViT model, I will compare its metrics with those of the 32-bit CNN model.
- **Record metrics for quantized 4-bit ViT for CNN comparison if feasible**
    - If time allows and I can achieve a 4-bit ViT model, I will compare its metrics with those of the 32-bit CNN model.

Date of Meetings:

- 03/19/26

Client Feedback on Milestone 5:
See Faculty Advisor Feedback below.

Faculty Advisor Feedback on Milestone 5:

Faculty Advisor Signature: _____ Date: 30 March 2026