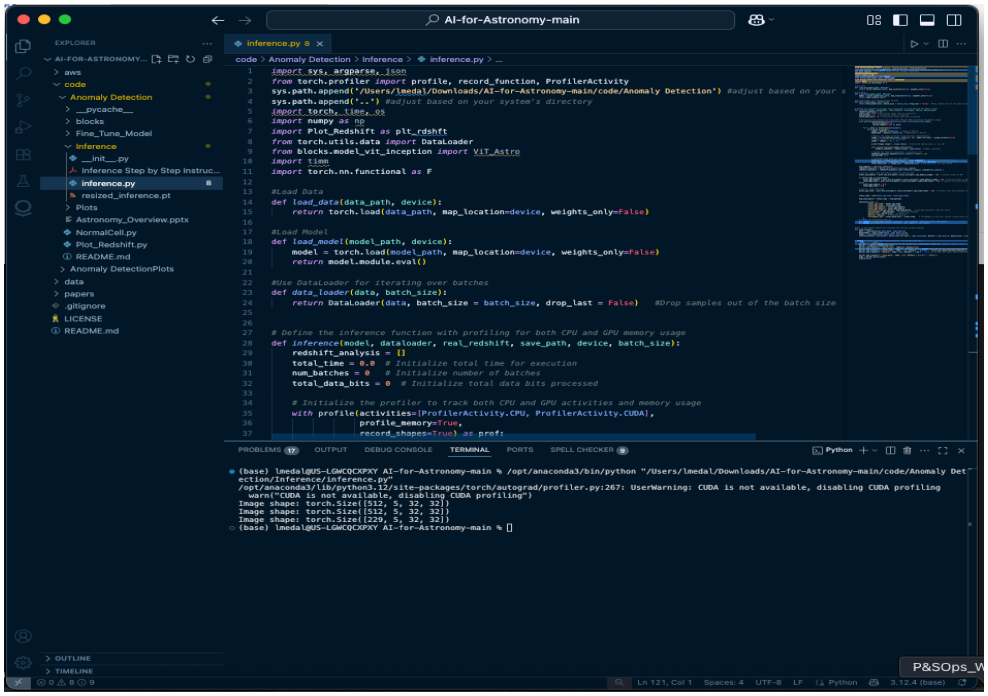
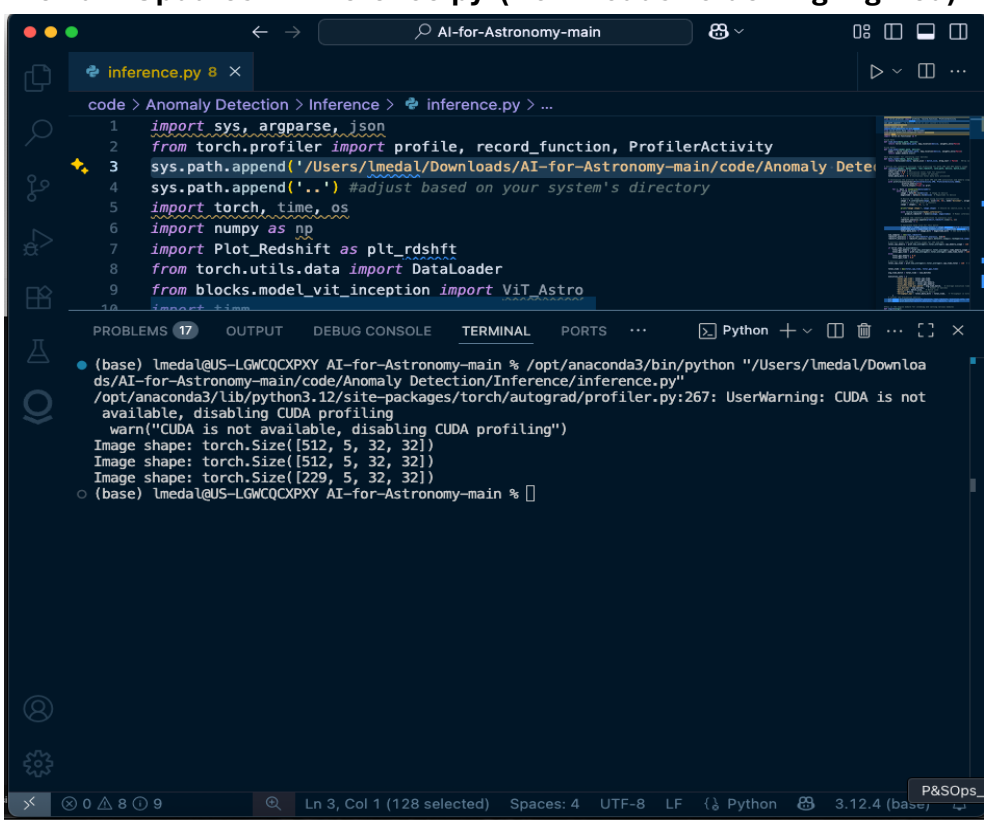
Step 3: Astronomy Inference

Team 5: Lionel Medal and Vicky Singh

Repository Cloning (Local VSCode)



File Path Updates in 'inference.py' (Downloads Folder Highlighted)



Documentation of Execution Time

The inference process was executed on a CPU using a batch size of 512. According to the Results.json output, the total CPU processing time across all batches was 9.56 seconds, with an average execution time per batch of 3.19 seconds. No GPU resources were utilized during this run, with the reported total GPU time being 0.0 seconds. The system achieved a throughput of approximately 21.5 million bits per second, demonstrating efficient performance for CPU-based execution under the given configuration.

Output Files

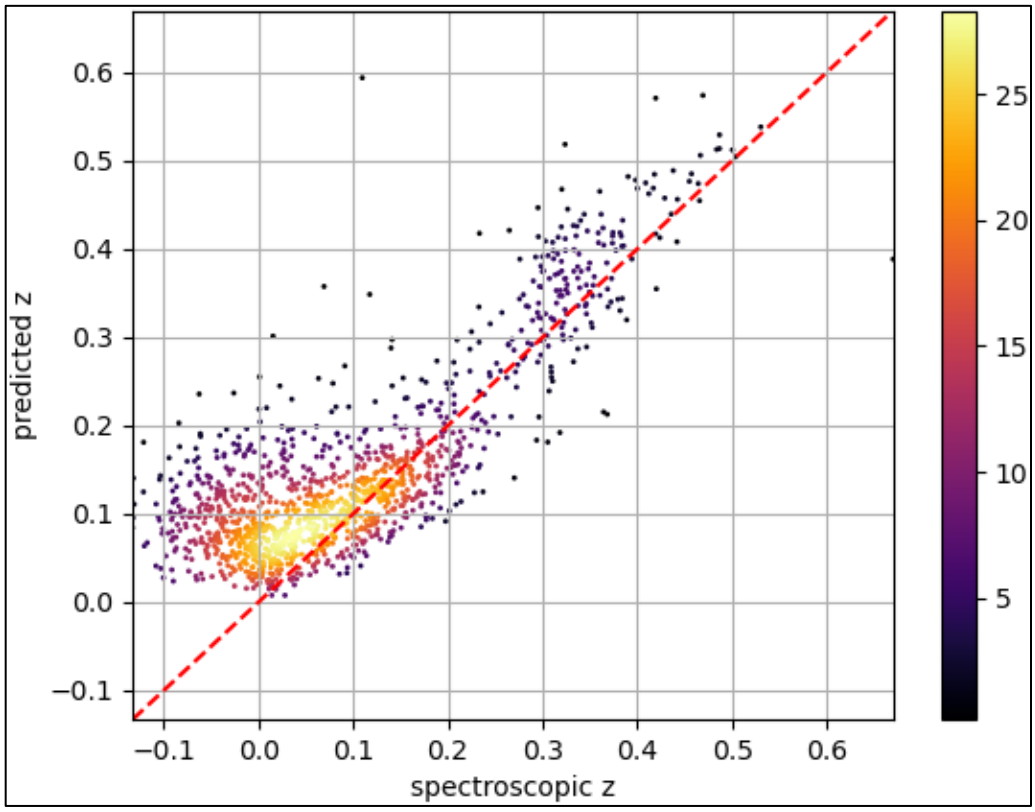


Figure 1. Inference Output File


```
{
  "total cpu time (second)": 9.558878423999927,
  "total gpu time (second)": 0.0,
  "execution time per batch (second)": 3.1862928079999757,
  "cpu memory (MB)": 31478.664708,
  "gpu memory (MB)": 0.0,
  "throughput(bps)": 21518474.33100082,
  "batch size": 512,
  "number of batches": 3,
  "device": "cpu",
  "MAE": 0.06971974536503017,
  "MSE": 0.008599726562012888,
  "Bias": -0.04611356450573312,
  "Precision": 0.061842061728239056,
  "R2": 0.26737546920776367
}
```


Figure 2. Results JSON Output File

Analysis of Inference Performance

- **Mean Absolute Error (MAE):** 0.0697
  Indicates the average absolute deviation between predicted and actual redshift values.
- **Mean Squared Error (MSE):** 0.0086
  Measures the average of the squared errors, placing greater emphasis on larger deviations.

- **Bias:** -0.046
  Reflects a slight tendency of the model to underpredict redshift values.
- **Precision (NMAD):** 0.0618
  Suggests a moderate spread in the prediction errors, based on the normalized median absolute deviation.
- **R^2 Score:** 0.26
  Indicates that approximately 26.7% of the variance in true redshift values is explained by the model. This relatively low score highlights room for further optimization or model refinement.

Deployment Options Comparison

Batch Size	Total Time (s)	Memory (MB)	Throughput (bps)	No. of Batches	Device
128	10.78	31,186.56	19,082,229	10	CPU
256	10.07	31,433.14	20,419,591	5	CPU
512	9.56	31,478.66	21,518,474	3	CPU
1024	9.87	31,651.86	20,834,646	2	CPU

Troubleshooting and Resolutions

Several issues were encountered and resolved during implementation:

- **Input Shape Mismatch**
  The model was configured to accept images with 5 channels and a resolution of 32×32, while the input data initially contained 64 channels and a different size. This was addressed by resizing the images to 32×32 and selecting only the first 5 channels for processing.
- **CUDA/GPU Profiling Errors**
  Running on a CPU-only system caused errors when the code attempted to access GPU attributes. The script was updated to check for CUDA availability and default to zero for GPU-related metrics when unavailable.
- **FileNotFoundError for Output Directory**
  The script initially failed when the specified output directory did not exist. This was resolved by adding logic to create the directory automatically before writing output files.
- **Path and Environment Adjustments**
  File paths within inference.py were modified to align with the local environment. Additionally, any missing Python dependencies were installed to ensure successful execution.