

Weekly Progress Report (Week of February 20–26, 2026)

Our project code is maintained on GitHub at the following repository:

https://github.com/zhengmiao1/ECS251_GPU.git

Project: ECS251 GPU Scheduling for Mixed LLM/VLM Workloads

Team: Zheng Miao, Zaishuo Xia, Qiyao Ma

1. Progress This Week

Zheng Miao

- Conducted systematic **parameter sensitivity analysis** on `aging_window` (60–300 s) and `short_threshold` (30–120 s) using the new `scripts/param_sweep.py`.
- Identified stable operating region: `aging_window` in [120, 240]s, `short_threshold` in [60, 90] s yield the best wait time/fairness trade-off.
- Formalized **final policy specification** with tuned defaults: `short_threshold` = 60.0 s, `aging_window` = 180.0 s.

Zaishuo Xia

- Implemented **SQLite persistence layer** (`scripts/db_store.py`) that stores every scheduling run, admission decision, and task result.
- Updated `scripts/experiment.py` to support a `--batch` flag for running all workload modes (mixed, `llm_heavy`, `vlm_heavy`) in a single invocation.
- Added `--out_db` option to route experiment results directly into the SQLite store.

Qiyao Ma

- Ran broader experiments: 5 seeds \times 3 workload modes \times 2 policies = 30 simulation runs (6,000 tasks per configuration).
 - Implemented `scripts/plot_results.py` to automate visualization.
 - Generated comparative boxplots for Average Waiting Time (AWT) and GPU utilization across different policies.
-

2. Updated File Structure

File	Description
<code>scripts/param_sweep.py</code>	New script for automated grid search over policy parameters
<code>scripts/db_store.py</code>	SQLite schema and insertion logic for experiment tracking
<code>scripts/plot_results.py</code>	Matplotlib and markdown-table generator from CSV outputs
<code>scripts/experiment.py</code>	Extended with <code>--batch</code> and <code>--out_db</code> flags

3. Plan for Next Week

Zheng Miao

- Write the scheduling policy section of the final report (formal specification + pseudocode)
- investigate priority-inversion edge cases.

Zaishuo Xia

- Extend db_store.py with analytical queries (wait distribution, GPU utilization timeline)
- add a --replay mode to simulate.py.

Qiyao Ma

- Finalize all figures with proper axis labels
- draft the evaluation section of the report with quantitative results tables.