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

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 × 3 workload modes × 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
scripts/param_sweep.py	New script for automated grid search over policy parameters
scripts/db_store.py	SQLite schema and insertion logic for experiment tracking
scripts/plot_results.py	Matplotlib and markdown-table generator from CSV outputs
scripts/experiment.py	Extended with --batch and --out_db flags