Weekly Progress Report (Week of February 13–19, 2026)

Project: ECS251 GPU Scheduling for Mixed LLM/VLM Workloads
Team: Zheng Miao, Zaishuo Xia, Qiyao Ma

1. Progress This Week

- **Zheng Miao**
  - Finalized practical scheduling policy components for integration:
    - Memory-feasible admission control
    - Short-task preference with anti-starvation aging
    - Explicit rejection rule for tasks exceeding GPU capacity
- **Zaishuo Xia**
  - Implemented a runnable scheduler prototype with two policies:
    - `memory` (memory-aware policy)
    - `fifo` (baseline)
  - Added structured JSONL event logging for scheduler decisions:
    - `dispatch`, `defer`, and `reject`
  - Extended simulation CLI to support:
    - Policy selection (`memory`, `fifo`, `both`)
    - Workload modes (`mixed`, `llm_heavy`, `vlm_heavy`)
    - Log output directory for trace collection
- **Qiyao Ma**
  - Defined and integrated evaluation metrics in code:
    - `completed_tasks`
    - `avg_wait_time`
    - `p95_wait_time`
    - `avg_turnaround`
    - `throughput`
    - `utilization`
    - `fairness_wait_std`
    - `oom_events`
  - Built a multi-seed experiment script to compare baseline vs. proposed policy and export report-ready results.

2. Code Deliverables Completed

- `scripts/scheduler.py`
  - Memory-aware scheduler improvements
  - FIFO baseline implementation
  - Anti-starvation aging logic
- `scripts/event_logger.py`
  - JSONL event logger for decision traces
- `scripts/simulate.py`

    - Mixed-workload generation and policy comparison in one run
- `scripts/metrics.py`
    - Expanded metric suite with corrected utilization computation
- `scripts/experiment.py`
    - Multi-seed baseline experiment runner
- `README.md`
    - Updated runnable commands and project structure documentation

3. Preliminary Outcome

- The prototype now supports direct, reproducible **policy comparison** under mixed workloads.
- Logging provides concrete evidence of scheduler behavior for analysis and reporting.
- Multi-seed experiments are producing stable summary tables for the upcoming evaluation section.

4. Plan for Next Week (February 20–26, 2026)

- **Zheng Miao**
    - Tune policy parameters (e.g., aging window, short-task threshold) and formalize final policy spec.
- **Zaishuo Xia**
    - Add persistence/DB integration path from scheduling decisions to stored records.
    - Improve experiment automation (CSV outputs + configurable batch runs).
- **Qiyao Ma**
    - Run broader experiments across workload mixes and resource settings.
    - Prepare comparative plots/tables (FIFO vs memory-aware) for the report draft.

5. Current Risks / Open Items

- Need real workload traces (or calibrated synthetic traces) to strengthen external validity.
- Parameter sensitivity analysis is still limited and should be expanded before final claims.