

PR #34181 完整报告

sgl-project/sglang

Fix mlx unit test batch mock

合并时间: 2026-08-10 00:55

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34181>

执行摘要

- 一句话: 修复 MLX 单测 mock 缺 mamba 字段导致的 CI 失败
- 推荐动作: 作为测试修复, 本 PR 适合快速浏览, 不值得精读。值得关注的点是: 生产代码新增字段时, 测试 mock 需要同步维护; 可结合 #33477 的 Mamba mask 复用逻辑理解该字段的来源。

功能与动机

PR body 指出, #33477 在 `_handle_finish_state_updated_req` 中新增了 `batch.mamba_track_mask_cpu` 访问, 但 MLX 单测 `test_finished_request_snapshots_before_release` 用裸 `SimpleNamespace()` 构造 `batch`, 导致 `AttributeError: 'types.SimpleNamespace' object has no attribute 'mamba_track_mask_cpu'`, `stage-a-unit-test-mlx` 在 main 上变红。因此本 PR 为 mock batch 补充三个 mamba 字段, 默认值与 `ScheduleBatch` 一致 (`None`), 让 mamba 边界分支短路。

实现拆解

1. 在 `test/registered/unit/hardware_backend/mlx/test_attention_patching.py` 的 `test_finished_request_snapshots_before_release` 中, 将 `batch = SimpleNamespace()` 改为显式携带三个字段: `mamba_track_mask_cpu`、`mamba_track_mask_next_cpu`、`mamba_decode_batch_idx_cpu`, 默认值均为 `None`。
2. 这三个字段与 `ScheduleBatch` 的构造默认值一致, 使 `_handle_finish_state_updated_req` 的 mamba 边界分支条件因 `None` 而短路, 不会进入需要真实张量的路径。
3. 测试原有的 `noop_stubs` 打桩、事件断言与 `finally` 恢复逻辑均未改动, 仅校正 mock 的形状。
4. 配套验证: 通过 issue 评论 `/rerun-test` 触发指定的 MLX 单测, GitHub Actions bot 确认 `ubuntu-latest` 上的该测试通过 (1 个测试)。

关键文件:

- `test/registered/unit/hardware_backend/mlx/test_attention_patching.py` (模块 MLX 测试; 类别 test; 类型 test-coverage): 唯一变更文件。修复 `test_finished_request_snapshots_before_release` 中 mock batch 缺少 mamba 追踪字段导致的 `AttributeError`, 使 `stage-a-unit-test-mlx` CI 恢复绿色。

关键符号: `test_finished_request_snapshots_before_release`,
`_handle_finish_state_updated_req`

关键源码片段

`test/registered/unit/hardware_backend/mlx/test_attention_patching.py`

唯一变更文件。修复 `test_finished_request_snapshots_before_release` 中 mock batch 缺少 mamba 追踪字段导致的 `AttributeError`, 使 `stage-a-unit-test-mlx` CI 恢复绿色。

```
# 修复前 batch 是空的 SimpleNamespace(), 而 _handle_finish_state_updated_req
# 会直接读取 batch.mamba_track_mask_cpu 等字段, 导致 AttributeError;
# 现在按 ScheduleBatch 的默认值 None 补齐三个 mamba 追踪字段,
# 使 mamba 边界分支因 None 短路, 测试事件序列保持预期。
batch = SimpleNamespace(
    mamba_track_mask_cpu=None, # CPU 侧 Mamba 追踪 mask
    mamba_track_mask_next_cpu=None, # 下一个 step 的追踪 mask
    mamba_decode_batch_idx_cpu=None, # decode 批次索引
)

# 调用被测试的生产方法, 此时 mamba 分支直接跳过
SchedulerBatchResultProcessor._handle_finish_state_updated_req(
    processor, req, batch, result, i, logits_output
)
```

评论区精华

本 PR 没有 review 评论。唯一互动是作者通过 issue 评论 `/rerun-test`
`test/registered/unit/hardware_backend/mlx/test_attention_patching.py` 触发重跑,
GitHub Actions bot 报告 `ubuntu-latest` 上的该测试通过。没有设计争论或未解决疑虑。

- 暂无高价值评论线程

风险与影响

- 风险: 改动仅涉及测试 mock, 无生产逻辑变更, 不存在回归、性能、安全影响。潜在风险在于 mock 字段手工维护, 若未来 `ScheduleBatch` 再新增 mamba 相关字段或改变默认值, 该测试可能再次抛 `AttributeError` 或与真实行为不一致。该风险仅影响 MLX 单元测试, 不扩散到运行时。
- 影响: 影响范围局限在 `test/registered/unit/hardware_backend/mlx/test_attention_patching.py` 与 `stage-a-unit-test-mlx` CI 步骤; 对最终用户和运行时无影响。对团队而言, 恢复 CI 绿色并解除因测试失败造成的合并阻塞, 影响程度低。
- 风险标记: mock 与真实 `ScheduleBatch` 可能漂移, 仅覆盖 MLX 单测

关联脉络

- PR #33477 [srt] Reuse batched Mamba boundary mask: 本 PR 是 #33477 引入的测试回归修复。#33477 在 `batch_result_processor.py` 的

_handle_finish_state_updated_req 中新增 batch.mamba_track_mask_cpu 访问，造成 MLX 单测 mock batch 缺失该字段而失败。