

PR #32448 完整报告

sgl-project/sglang

[MLX] Move fused swiglu tests to test/registered so CI collects them

合并时间: 2026-07-29 10:04

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32448>

执行摘要

- 一句话: 搬迁 MLX fused-swiglu 测试到 CI 可见目录
- 推荐动作: 值得合并, 修复 CI 死代码问题, 提升测试覆盖。建议未来迁移同类包内测试时参考此模式。

功能与动机

该测试文件原先位于 `python/sglang/srt/hardware_backend/mlx/moe/tests/`, 由于 `run_suite.py` 只收集 `test/registered/**` 且 `check-no-registered-tests-in-package` 钩子禁止包内注册, 该测试在 CI 中完全不可见, 成为死代码。

实现拆解

1. 文件搬迁: 将 `test_fused_swiglu.py` 从 `python/sglang/srt/hardware_backend/mlx/moe/tests/` 移至 `test/registered/unit/hardware_backend/mlx/`, 并删除原空包目录 `../moe/tests/__init__.py`。
2. CI 注册: 添加 `register_cpu_ci(est_time=1, suite="base-a-test-cpu")` 和 `register_mlx_ci(est_time=45, suite="stage-a-unit-test-mlx")` 引入注册接口。
3. 导入守卫重构: 将顶层 `mx = pytest.importorskip("mlx.core")` 替换为 `importlib.util.find_spec` 守卫 + `pytestmark` 模块级跳过, 解决 CI 以 `python3 file.py` 方式执行时 `importorskip` 导致非零退出的问题。同时增加平台检查 `Darwin/arm64`。
4. 入口添加: 在文件末尾添加 `if __name__ == "__main__": sys.exit(pytest.main([__file__, "-v"]))`, 满足 `collect_tests` 要求。

关键文件:

- `test/registered/unit/hardware_backend/mlx/test_fused_swiglu.py` (模块 测试; 类别 `test`; 类型 `rename-or-move`): 核心变更文件: 测试从包内搬迁到 CI 可见目录, 并重构导入守卫与注册逻辑。
- `python/sglang/srt/hardware_backend/mlx/moe/tests/__init__.py` (模块 测试; 类别 `test`; 类型 `deletion`): 空包目录删除, 配合文件搬迁。

关键符号: 未识别

关键源码片段

test/registered/unit/hardware_backend/mlx/test_fused_swiglu.py

核心变更文件：测试从包内搬迁到 CI 可见目录，并重构导入守卫与注册逻辑。

```
"""Numerical equivalence and eligibility tests for the Path B fused swiglu kernel.
```

Two groups:

- * Model-based equivalence (@requires_model): loads a small MoE model, runs the fused gate_qmv + silu + x_up kernel against the unfused reference (mx.gather_qmm + nn.silu(gate) * x_up) on both the unsorted and sorted paths. Gated by SGLANG_MLX_TEST_MODEL so CI hosts without a model cache skip them (stage-a sets HF_HUB_OFFLINE=1, so they stay skipped there).
- * Synthetic eligibility (no model, MLX only): the learned-bias fallback. The fused kernel recomputes the gate matmul and has no slot for the per-expert learned bias QuantizedSwitchLinear adds after the matmul, so can_fuse must exclude a gate carrying one, and the patch must leave such a layer unfused. These run wherever MLX is available (Apple Silicon).

Registered on the CPU suite but skipped wherever mlx is absent; runs for real only on Apple Silicon via stage-a-unit-test-mlx.

```
"""
```

```
import importlib.util
import os
import platform
import sys
```

```
import pytest
```

```
from sglang.test.ci.ci_register import register_cpu_ci, register_mlx_ci
```

```
# Register the file in two CI suites:
```

```
# - CPU suite: always runs but skipped on non-Apple-Silicon hosts.
```

```
# - MLX suite: runs only on Apple Silicon with MLX installed.
```

```
register_cpu_ci(est_time=1, suite="base-a-test-cpu")
```

```
register_mlx_ci(est_time=45, suite="stage-a-unit-test-mlx")
```

```
# Platform and dependency guards: we use find_spec instead of importorskip
```

```
# because the CI runner executes registered files as 'python3 file.py';
```

```
# a top-level Skipped from importorskip would escape uncaught and exit
```

```
# non-zero on MLX-less hosts before pytest.main() ever runs.
```

```
_IS_APPLE_SILICON = platform.system() == "Darwin" and platform.machine() == "arm64"
```

```
_HAS_MLX = (
```

```
    importlib.util.find_spec("mlx") is not None
```

```
    and importlib.util.find_spec("mlx_lm") is not None
```

```
)
```

```
pytestmark = pytest.mark.skipif(
```

```
    not (_IS_APPLE_SILICON and _HAS_MLX),
```

```
    reason="Apple-Silicon-only test (requires Darwin/arm64 + mlx + mlx_lm)",
```

```

)

if _HAS_MLX:
    import mlx.core as mx

# Model-based tests need a real checkpoint; synthetic tests below do not.
requires_model = pytest.mark.skipif(
    not os.environ.get("SGLANG_MLX_TEST_MODEL"),
    reason="Set SGLANG_MLX_TEST_MODEL to a HuggingFace model id to enable",
)

# ... (rest of test functions remain unchanged) ...

if __name__ == "__main__":
    sys.exit(pytest.main([__file__, "-v"]))

```

评论区精华

无实质讨论；审核人 yeahdongcn 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅涉及测试文件搬迁和导入守卫调整，无任何运行时或生产代码改动。但需确保新路径下的导入路径正确（已确认），且 register_mlx_ci 的 est_time=45 不应导致 CI 超时（实测约 43s）。
- 影响：影响范围限于 CI 流程和 Apple Silicon 开发者。原先不可见的 11 个测试用例（含 1 个关键数值正确性检查）将自动在 MLX 目标 CI 上运行，提升 fused-swiglu 内核的质量保障。在无 MLX 的环境（如 CPU 注册）中测试会被正确跳过，无副作用。
- 风险标记：测试搬迁，无运行时影响

关联脉络

- PR #32447 [MLX] Fix overlap-loop request bookkeeping and graceful shutdown: 同一作者和审核人，均为 MLX 相关修复