

PR #36639 完整报告

sgl-project/sglang

[CI] Fix Q8KV8 sparse prefill test fixture

合并时间: 2026-08-27 16:00

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36639>

执行摘要

- 一句话: 修复 Q8KV8 稀疏预填充测试夹具
- 推荐动作: 值得快速浏览, 了解测试夹具与后端重构的耦合关系。建议关注后续是否还有其他测试需要类似调整。

功能与动机

PR body 指出, 需要保持 Q8KV8 稀疏预填充夹具与 #35947 缓存构建器重构的兼容性。由于 #35947 改变了后端中缓存池的构建方式, 测试夹具中未初始化的 `token_to_kv_pool` 导致测试失败。

实现拆解

1. 定位问题: 在 `test_q8kv8_sparse_prefill_backend.py` 的 `_make_sparse_prefill_case` 函数中, 测试创建了 `token_to_kv_pool` 对象但未将其赋值给后端。
2. 修复方式: 在 `token_to_kv_pool` 创建后, 立即添加 `backend.token_to_kv_pool = token_to_kv_pool` 赋值语句。
3. 验证: 通过运行测试 `test_q8kv8_sparse_prefill_backend.py` 确认修复有效, 测试通过。

关键文件:

- `test/registered/kernels/ops/attention/test_q8kv8_sparse_prefill_backend.py` (模块 测试; 类别 test; 类型 test-coverage): 修复测试夹具, 将 `token_to_kv_pool` 赋值给后端, 以适配 #35947 的重构。

关键符号: `_make_sparse_prefill_case`

关键源码片段

`test/registered/kernels/ops/attention/test_q8kv8_sparse_prefill_backend.py`

修复测试夹具, 将 `token_to_kv_pool` 赋值给后端, 以适配 #35947 的重构。

```
# 测试夹具中初始化后端池引用
from types import SimpleNamespace
import torch
```

```
# 创建后端后, 将 token_to_kv_pool 赋值给后端, 以便后续逻辑使用
```

```
token_to_kv_pool = _TokenToKVPool(  
    swa_key_buffer=quant_k_cache,  
    extra_key_buffer=extra_k_cache,  
    full_to_swa_index_mapping=torch.arange(  
        total_slots, dtype=torch.int64, device=device  
    ),  
    page_size=page_size,  
)  
backend.token_to_kv_pool = token_to_kv_pool
```

评论区精华

Review 讨论较少，作者请求合并以解锁 CI，审查者 mmangkad 批准了变更。

- 暂无高价值评论线程

风险与影响

- 风险：该变更仅涉及测试夹具，不影响生产代码。风险极低，但应在 CI 中确保所有后端相关测试通过，防止回归。
- 影响：影响范围仅限于 Q8KV8 稀疏预填充测试，使 CI 恢复通过。对用户和系统无直接影响。
- 风险标记：测试仅适配，无生产代码变更

关联脉络

- PR #35947 Publish gated DSV4 DFLASH-family target-prefill read completion: 本 PR 是针对 #35947 重构的测试修复。