PR #32047 完整报告

sgl-project/sglang

[CI] Fix stale per-token group quant callers

合并时间：2026-07-22 19:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32047

执行摘要

- 一句话：修复过时的 per-token group quant 调用者
- 推荐动作：建议快速合并。该 PR 修复了 CI 中的大量测试失败，是必要的清理工作。改动简单明了，审核已通过。无需精读，但可关注类似 PR 中因参数废弃导致的调用者更新。

功能与动机

PR body 明确指出：'Caused by #30924:

- 1,760 failed tests in the H100 sgl-kernel-unit-test job. - 1,982 failed tests in the B200 sgl-kernel-b200-test job.' 所有失败均为 test_per_token_group_quant_8bit.py 中的参数化测试。

实现拆解

1. 基准测试文件 (sgl-kernel/benchmark/bench_per_token_group_quant_8bit.py)：
   - 移除未使用的导入：time、partial、Path、create_per_token_group_quant_fp8_output_scale。
   - 在 benchmark 函数中，将 sglang provider 的调用从 partial(sglang_per_token_group_quant_8bit, enable_v2=True) 改为直接使用 sglang_per_token_group_quant_8bit，因为 enable_v2 参数已被移除。
   - 将 bench_fn 从 lambda 改为嵌套函数定义，逻辑不变。
2. 测试文件 (sgl-kernel/tests/test_per_token_group_quant_8bit.py)：
   - 移除未使用的导入：os、time、Path、get_bool_env_var。
   - 在调用 sglang_per_token_group_quant_8bit 时移除 enable_v2=True 参数，与基准测试文件中的变更一致。

关键文件：

- sgl-kernel/benchmark/bench_per_token_group_quant_8bit.py（模块 基准测试；类别 source；类型 core-logic；符号 bench_fn）：核心变更文件，移除了 partial 和 enable_v2=True，并清理了未使用的导入。
- sgl-kernel/tests/test_per_token_group_quant_8bit.py（模块 测试；类别 test；类型 test-coverage）：测试文件，移除了 enable_v2=True 参数和未使用的导入，使测试与 API 保持一致。

关键符号：bench_fn

关键源码片段

sgl-kernel/benchmark/bench_per_token_group_quant_8bit.py

核心变更文件，移除了 partial 和 enable_v2=True，并清理了未使用的导入。

```
def benchmark(
    num_tokens, hidden_dim, group_size, num_ranks, dst_dtype, flags, provider
):
    # ... 数据准备 ...
    fn, kernel_names = {
        "triton": (
            triton_per_token_group_quant_8bit,
            "_per_token_group_quant_8bitl_silu_and_mul_post_quant_kernel",
        ),
        "sglang": (
            sglang_per_token_group_quant_8bit, # 直接引用，不再使用 partial 传递 enable_v2=True
            "per_token_group_quant_8bit_kernel",
        ),
    }[provider]

    def bench_fn(): # 从 lambda 改为嵌套函数，逻辑不变
        return fn(
            x=x,
            masked_m=masked_m,
            group_size=group_size,
            dst_dtype=dst_dtype,
            **{k: v for k, v in flags.items() if k not in ["masked_layout_mode"]},
        )

    time_s = bench_kineto(
        bench_fn, kernel_names=kernel_names, num_tests=300 if mode_concentrated else 30
    )
    return time_s * 1e6
```

sgl-kernel/tests/test_per_token_group_quant_8bit.py

测试文件，移除了 enable_v2=True 参数和未使用的导入，使测试与 API 保持一致。

```
# 生成参考结果（Triton 实现）
x_q_triton, x_s_triton = _postprocess(
    *triton_per_token_group_quant_8bit(**execute_kwargs)
)
# 生成待测结果（SGL kernel 实现），不再传递已移除的 enable_v2 参数
x_q_sglang, x_s_sglang = _postprocess(
    *sglang_per_token_group_quant_8bit(**execute_kwargs)
)

try:
```

```
        assert_all_close_or_tiny_diff(x_q_triton, x_q_sglang)
        torch.testing.assert_close(
            x_s_triton.contiguous(),
            x_s_sglang.contiguous(),
            rtol=1e-3,
            atol=1e-5,
            msg=lambda message: message + f" {x_s_triton=} {x_s_sglang=}",
        )
    except AssertionError:
        # 详细打印调试信息
        # ...
        raise
```

评论区精华

该 PR 涉及 2 个 review 评论，但内容为空，无实质性讨论。审核人 BBuf 已批准该 PR，无合并冲突。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅限于移除已废弃的参数和未使用的导入，逻辑行为未变。所有测试在合并前已通过（PR 修复了测试失败），因此无回归风险。无性能、安全或兼容性影响。
- 影响：影响范围有限，仅涉及 sgl-kernel 模块的基准测试和测试代码。修复了 CI 中约 3,742 个测试用例，确保量化内核的验证流水线恢复稳定。对用户无直接影响。
- 风险标记：暂无

关联脉络

- PR #30924 [Kernel] Make enable_v2 default True in per_token_group_quant_8bit: 本 PR 修复了由 #30924 引入的调用者不兼容问题。