

PR #29636 完整报告

sgl-project/sglang

[Kernel] Strengthen kernel shape coverage

合并时间: 2026-07-01 15:44

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29636>

执行摘要

- 一句话: 加强 JIT 和 sgl-kernel 测试的形状覆盖
- 推荐动作: 值得阅读的 PR。它展示了如何为 kernel 测试建立高效的生产环境覆盖策略——通过 `get_ci_test_range` 分离 CI 和本地 /nightly 测试范围。与 RFC #29630 的关系进一步说明了测试基础设施对架构演进的重要性。建议所有 kernel 贡献者学习此模式, 在新增 kernel 时同步添加 production-like shape 测试。

功能与动机

PR body 明确说明 'Part of the test-hardening baseline for RFC #29630' 以及 'This also gives future agent-kernel work a clearer correctness baseline before kernel APIs move around.' 主要动机是为即将进行的统一 kernel 命名空间重构建立坚实的测试覆盖基础, 确保核心 kernel 在接近生产的输入范围内正确工作。

实现拆解

1. 在 `sgl-kernel/tests/test_norm.py` 中新增 `assert_close_norm` 辅助函数和 `PRODUCTION_LIKE_NORM_CASES`、`PRODUCTION_LIKE_FUSED_ADD_RMSNORM_CASES` 两个 shape 表, 并分别添加 `test_norm_production_like_shapes` 和 `test_fused_add_rmsnorm_production_like_shapes` 测试函数, 覆盖 `bfloat16` 下典型 batch 和 hidden 维度组合 (如 `38x4096`、`1240x1536`、`7807x128` 等)。
2. 在多个 JIT kernel 测试文件 (如 `test_per_token_group_quant_8bit.py`、`test_per_token_group_quant_8bit_v2.py`、`test_dsv3_router_gemm.py`、`test_rmsnorm.py` 等) 中, 将原本使用 `@pytest.mark.parametrize` 全量组合的测试参数改为通过 `get_ci_test_range` 生成。该函数接受全量组合列表和一组代表性子集, 在 CI 中只运行代表性子集, 而在本地或 nightly 中运行全量组合, 从而平衡 CI 速度与覆盖深度。
3. 在 `sgl-kernel/tests/test_fp8_gemm.py` 中新增 `PRODUCTION_LIKE_FP8_GEMM_CASES` 和 `test_accuracy_production_like_shapes` 测试, 使用与生产场景相似的 M、N、K 维度 (如 `189x4608x8192`) 验证 FP8 GEMM 精度。类似地, 在 `sgl-kernel/tests/test_per_token_quant_fp8.py` 中添加了代表性 shape 覆盖。
4. 在 `test_per_token_group_quant_8bit_v2.py` 中调整了 `sgl_kernel` 导入逻辑: 当 AOT 参考 op 缺失时, 如果 CUDA 可用则抛出 `ImportError` (在 CI 中确保失败), 否则跳过整个模块。这一改动引发了一条 review 评论。

5. 同步调整了 `bench_sparse_mla_q8kv8_prefill_sm90.py` 中的 `lambda` 函数为具名内部函数，便于 profiling 工具识别。

关键文件：

- `sgl-kernel/tests/test_norm.py`（模块 归一化；类别 test；类型 test-coverage；符号 `assert_close_norm`, `test_norm_production_like_shapes`, `test_fused_add_rmsnorm_production_like_shapes`）：新增了 production-like shape 测试和柔性断言辅助函数，是本次测试增强的核心文件之一。
- `test/registered/jit/test_per_token_group_quant_8bit.py`（模块 Token 分组量化；类别 test；类型 test-coverage）：改为使用 `get_ci_test_range` 控制 CI 测试范围，并重构了 `flags` 定义，是测试覆盖策略变更的代表。
- `test/registered/jit/test_per_token_group_quant_8bit_v2.py`（模块 V2 分组量化；类别 test；类型 test-coverage）：修改了 AOT 参考 op 的导入逻辑，并同样引入 `get_ci_test_range`，引发了重要的 review 讨论。
- `test/registered/jit/test_dsv3_router_gemm.py`（模块 路由矩阵；类别 test；类型 test-coverage）：使用 `get_ci_test_range` 简化和缩小 CI 测试网格，是路由 GEMM kernel 的测试增强。

关键符号：`assert_close_norm`, `test_norm_production_like_shapes`,
`test_fused_add_rmsnorm_production_like_shapes`, `test_accuracy_production_like_shapes`

关键源码片段

`sgl-kernel/tests/test_norm.py`

新增了 production-like shape 测试和柔性断言辅助函数，是本次测试增强的核心文件之一。

```
def assert_close_norm(actual, expected, dtype):
    # 对 bfloat16 使用更宽松的容忍度，因为其精度较低
    if dtype is torch.bfloat16:
        torch.testing.assert_close(actual, expected, rtol=1e-2, atol=2e-2)
    else:
        torch.testing.assert_close(actual, expected, rtol=1e-3, atol=1e-3)
```

```
# 代表生产场景的 Norm 输入维度：(batch_size, hidden_size, dtype)
PRODUCTION_LIKE_NORM_CASES = [
    (38, 4096, torch.bfloat16), # 典型推理 batch + LLM hidden
    (1240, 1536, torch.bfloat16), # 较大 batch, 中等 hidden
    (7807, 128, torch.bfloat16), # 极大 batch, 小 hidden (embedding 等)
]
```

```
@pytest.mark.parametrize("batch_size,hidden_size,dtype", PRODUCTION_LIKE_NORM_CASES)
def test_norm_production_like_shapes(batch_size, hidden_size, dtype):
    x = torch.randn(batch_size, hidden_size, dtype=dtype, device="cuda")
    w = torch.randn(hidden_size, dtype=dtype, device="cuda")
    y_ref = llama_rms_norm(x, w)
    enable_pdl = is_arch_support_pdl()
    y = sgl_kernel.rmsnorm(x, w, enable_pdl=enable_pdl)
```

```

assert_close_norm(y_ref, y, dtype)

# 代表生产场景的 fused-add-rmsnorm 输入维度
PRODUCTION_LIKE_FUSED_ADD_RMSNORM_CASES = [
    (39, 4096, torch.bfloat16),
    (39, 8192, torch.bfloat16),
    (89, 4096, torch.bfloat16),
]

@pytest.mark.parametrize(
    "batch_size,hidden_size,dtype", PRODUCTION_LIKE_FUSED_ADD_RMSNORM_CASES
)
def test_fused_add_rmsnorm_production_like_shapes(batch_size, hidden_size, dtype):
    eps = 1e-6
    x = torch.randn(batch_size, hidden_size, dtype=dtype, device="cuda")
    residual = torch.randn_like(x)
    weight = torch.randn(hidden_size, dtype=dtype, device="cuda")
    x_native, residual_native = fused_add_rms_norm(
        x.clone(), residual.clone(), weight, eps
    )
    x_fused = x.clone()
    residual_fused = residual.clone()
    enable_pdl = is_arch_support_pdl()
    sgl_kernel.fused_add_rmsnorm(
        x_fused, residual_fused, weight, eps, enable_pdl=enable_pdl
    )
    assert_close_norm(x_fused, x_native, dtype)
    torch.testing.assert_close(residual_fused, residual_native, rtol=1e-3, atol=1e-3)

```

评论区精华

在 `test/registered/jit/test_per_token_group_quant_8bit_v2.py` 中，`gemini-code-assist[bot]` 指出模块级别的 `ImportError`（当 `sgl_kernel` 缺失但 CUDA 可用时）会破坏本地 `pytest` 收集阶段，导致开发者无法运行其他测试。建议使用 `is_in_ci()` 判断：在 CI 中报错，在本地跳过。但 PR 最终合并时仍保留了 `raise ImportError`，未采纳该建议，可能因为该测试文件强依赖 `sgl_kernel`，且 CI 环境必然可用。这是 PR 中唯一的设计争议。

- `test_per_token_group_quant_8bit_v2.py` 中 `ImportError` 应区分 CI 与本地环境 (correctness): PR 作者未修改，最终代码保留了 `raise ImportError`。可能因为该测试强依赖 `sgl_kernel`，且 CI 中必然可用。

风险与影响

- 风险：
 - 测试覆盖风险：`get_ci_test_range` 在 CI 中只运行小规模 frozen 表，可能遗漏某些边界形状。但 nightly 测试保留全量覆盖，且代表性子集由作者根据生产经验选取，风险可控。
 - 导入风险：`test_per_token_group_quant_8bit_v2.py` 中 `sgl_kernel` 缺失时直接抛出 `ImportError`，可能导致本地开发者无法执行该文件及其他测试（如果使用 `pytest --co`

收集)。但实际影响范围仅限于该文件，且 CI 中会正确报错。

- 回归风险：新增 production-like 测试可能对特定硬件（如 B200）有额外依赖，但 PR 作者已进行 B200 直接 kernel smoke 验证。
- 影响：
 - 对用户：无直接用户感知影响，均为测试基础设施改进。
 - 对 CI 系统：JIT kernel 测试的 CI 执行时间更稳定，覆盖更聚焦于生产相关形状，避免全量组合导致的超时。
 - 对开发者：新增测试使 kernel 重构（如 RFC #29630 的统一命名空间）有更可靠的正确性基线；但 `test_per_token_group_quant_8bit_v2.py` 的 `ImportError` 可能对本地无 `sgl_kernel` 的开发者造成不便。
 - 对团队：为后续 kernel 组织工作提供了验证框架，并可推广到更多 kernel 测试。
 - 风险标记：测试覆盖策略变更，`sgl_kernel` 导入依赖，CI 覆盖依赖 frozen 表

关联脉络

- PR #29630 [RFC] Introduce a unified `sglang.kernels` namespace for kernel organization and dispatch: 本 PR 是该 RFC 的测试加固基线的一部分，为统一 kernel 命名空间提供正确性保障。
- PR #29715 [CI] Migrate JIT tests missed by #29066 to `runner_config` registration: 该 PR 的提交历史中 cherry-pick 了 #29715，以合并 CI 注册变更。