

PR #32937 完整报告

sgl-project/sglang

[CI] Fix MoE compile and DSA indexer regressions

合并时间: 2026-07-31 06:21

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32937>

执行摘要

- 一句话: 修复 MoE 编译和 DSA indexer 回归问题
- 推荐动作: 值得精读: 变更虽小但展示了 CI 回归治理的优秀实践——作者未 force-merge, 而是回查失败测试并提交补丁, 而且通过 /rerun-test 验证修复。推荐开发者在重构数据类型时同步检查所有位置解构代码, 避免类似回归。

功能与动机

从 PR body 可知, PR#31888 合并前其 CI 跑 base 分支时已显示两个测试失败但被 force-merged: `test_torch_compile_moe.py` 报 `ValueError: too many values to unpack (expected 3)`, `TestDSAIndexer.test_forward_decode_mode` 报 `AttributeError: 'ServerArgs' object has no attribute 'enable_two_batch_overlap'`。这些回归会阻塞后续 CI 流水线。

实现拆解

1. 修复 MoE 字段访问方式: 在 `python/sglang/srt/layers/moe/fused_moe_native.py` 的 `fused_moe_forward_native` 函数中, 将 `x, x_scale, topk_output = dispatch_output` 改为 `x = dispatch_output.hidden_states; topk_output = dispatch_output.topk_output`。因为 `StandardDispatchOutput` 已扩充为四个字段 (新增 `x_scale`? 实际已从 `namedtuple` 变为 `dataclass` 等), 按位置解包会抛异常, 改为按属性访问更健壮。
2. 补充 DSA indexer 测试 mock 的缺失属性: 在 `test/registered/kernels/ops/attention/test_dsa_indexer.py` 的 `MockServerArgs` 字典中增加 `"enable_two_batch_overlap": False`。该属性是 DSA 后端新增的配置项, 测试中的 mock 未同步更新导致 `AttributeError`。

关键文件:

- `python/sglang/srt/layers/moe/fused_moe_native.py` (模块 MoE 调度; 类别 source; 类型 core-logic; 符号 `fused_moe_forward_native`): 修复 MoE torch.compile 路径中 `StandardDispatchOutput` 解包异常的核心改动。通过按属性访问代替位置解包, 避免因新增字段而崩溃。
- `test/registered/kernels/ops/attention/test_dsa_indexer.py` (模块 DSA 索引器; 类别 test; 类型 test-coverage): 修复 DSA indexer 测试因 `ServerArgs` mock 缺少新属性而崩溃的问题。增加 `enable_two_batch_overlap: False` 默认值。

关键符号: `fused_moe_forward_native`

关键源码片段

python/sglang/srt/layers/moe/fused_moe_native.py

修复 MoE torch.compile 路径中 StandardDispatchOutput 解包异常的核心改动。通过按属性访问代替位置解包，避免因新增字段而崩溃。

```
# 修复前：通过位置解包，当 StandardDispatchOutput 新增字段时会抛出
# ValueError: too many values to unpack (expected 3)
# x, x_scale, topk_output = dispatch_output

# 修复后：按明确的属性名取值，不依赖元组长度
x = dispatch_output.hidden_states
topk_output = dispatch_output.topk_output
# x_scale 在函数后续并未使用，因此直接忽略，避免未打包变量问题
```

test/registered/kernels/ops/attention/test_dsa_indexer.py

修复 DSA indexer 测试因 ServerArgs mock 缺少新属性而崩溃的问题。增加 `enable_two_batch_overlap: False` 默认值。

```
# mock 对象中增加 DSA 后端新增的属性，避免测试因 AttributeError 失败
self.server_args = type(
    "ServerArgs",
    (),
    {
        # ... 其他已有属性 ...
        "enable_two_batch_overlap": False, # 新增默认值
    },
)()
```

评论区精华

PR 没有 review 评论，作者 mmangkad 通过 `/rerun-test` 触发 CI 验证修复，两次测试均通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：两处修复均为局部且直接。MoE 字段访问改为按名取值，不会改变原有语义；测试 mock 增加字段默认值，不影响其他测试。但需要注意：如果未来 StandardDispatchOutput 结构再次变更，按属性访问仍可能因字段重命名而出错，但比位置解包更安全。
- 影响：
 - 影响范围：仅影响两个测试用例所在的模块：MoE 的 torch.compile 路径和 DSA indexer 测试。
 - 影响程度：修复后 CI 可正常通过，不会阻塞其他 PR 合并。
 - 用户感知：无。
 - 风险标记：低风险

关联脉络

- PR #31888 [DSA] Q8KV8 FP8 Sparse Prefill on GLM-5.2 & DeepSeek-V3.2: Q8-Path & Shared-Path Optimizations: 本 PR 修复了 #31888 合并后引入的回归。