PR #23973 完整报告

sgl-project/sglang

[Fix] FP8 Qwen3-Next quant error by removing fallback fused shards

合并时间：2026-04-30 05:33

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23973

执行摘要

- 一句话：修复 Qwen3-Next FP8 量化加载错误
- 推荐动作：该 PR 值得精读，因为它展示了 FP8 量化配置中 modules_to_not_convert 与融合层映射的微妙交互，以及回归问题的诊断过程。测试文件的设计清晰且具有代表性，可作为类似场景的测试模板。

功能与动机

修复由 #23467 引入的回归问题，该问题导致 Qwen3-Next FP8 在 B200 4-GPU 上服务器初始化时失败，报错："ValueError: Weight output_partition_size = 8 is not divisible by weight quantization block_n = 128"。PR body 详细描述了根因：in_proj_ba 和 in_proj_qkvz 是 Qwen3-Next FP8 checkpoint 中真实的统一张量（明确列在 modules_to_not_convert 中），但 #23467 将它们加入 _FALLBACK_FUSED_SHARDS 后，会被解构为不存在的子张量名（如 in_proj_b、in_proj_a），导致 is_layer_skipped 错误地返回 False，最终触发 Fp8LinearMethod.validate_block_quant_shapes 的校验失败。

实现拆解

实现分为两部分：

1. 源码修复（python/sglang/srt/layers/quantization/utils.py）：从 _FALLBACK_FUSED_SHARDS 字典中删除 "in_proj_ba": ["in_proj_b", "in_proj_a"] 和 "in_proj_qkvz": ["in_proj_qkv", "in_proj_z"] 两个条目。这个字典作为 packed_modules_mapping 缺失时的回退映射，用于将融合线性层名映射到其分片名。删除后，in_proj_ba 和 in_proj_qkvz 不再被视为虚拟融合名，因此 is_layer_skipped 会直接使用 modules_to_not_convert 中的原始名进行路径匹配，正确跳过这些层。
2. 新增回归测试（test/registered/quant/test_is_layer_skipped.py）：新增 CPU 回归测试文件（约 5s，stage-a-test-cpu），包含三个测试用例：
   - test_qwen3_next_in_proj_ba_is_skipped：验证 in_proj_ba 在 modules_to_not_convert 中时正确被跳过。
   - test_qwen3_next_in_proj_qkvz_is_skipped：验证 in_proj_qkvz 在 modules_to_not_convert 中时正确被跳过。
   - test_mlp_gate_does_not_match_gate_up_proj：验证 mlp.gate 不会错误匹配 mlp.gate_up_proj，确保 #23467 的原始动机仍被保留。

关键文件：

- python/sglang/srt/layers/quantization/utils.py（模块 量化层；类别 source；类型 core-logic；符号 _FALLBACK_FUSED_SHARDS, is_layer_skipped）：核心修复文件：删除 _FALLBACK_FUSED_SHARDS 中错误的两行，修复回归问题。
- test/registered/quant/test_is_layer_skipped.py（模块 测试；类别 test；类型 test-coverage；符号 _qwen3_next_ignored_layers, TestIsLayerSkipped, test_qwen3_next_in_proj_ba_is_skipped, test_qwen3_next_in_proj_qkvz_is_skipped）：新增回归测试文件，覆盖修复场景并确保 #23467 的原始动机仍被保留。

关键符号：is_layer_skipped, _module_path_match

关键源码片段

python/sglang/srt/layers/quantization/utils.py

核心修复文件：删除 _FALLBACK_FUSED_SHARDS 中错误的两行，修复回归问题。

```
# python/sglang/srt/layers/quantization/utils.py
# 关键片段：_FALLBACK_FUSED_SHARDS 的定义（修复后）

# Known fused-linear -> shard names. Used as a fallback when the quant
# config doesn't ship packed_modules_mapping (typical for HF FP8 configs).
# 注意：in_proj_ba 和 in_proj_qkvz 已被移除，因为它们是 Qwen3-Next FP8
# 中真实的统一张量（非虚拟融合名），不应被解构。
_FALLBACK_FUSED_SHARDS: Mapping[str, List[str]] = {
    "qkv_proj": ["q_proj", "k_proj", "v_proj"],
    "gate_up_proj": ["gate_proj", "up_proj"],
}
```

test/registered/quant/test_is_layer_skipped.py

新增回归测试文件，覆盖修复场景并确保 #23467 的原始动机仍被保留。

```
# test/registered/quant/test_is_layer_skipped.py
# 新增的回归测试文件

from sglang.srt.layers.quantization.utils import is_layer_skipped
from sglang.test.ci.ci_register import register_cpu_ci
from sglang.test.test_utils import CustomTestCase

register_cpu_ci(est_time=5, suite="stage-a-test-cpu")

# Qwen3-Next FP8 的 packed_modules_mapping 等价物（来自 qwen3_next.py:908-911）
# 注意：in_proj_ba 和 in_proj_qkvz 被故意省略，因为它们是真实统一张量
QWEN3_NEXT_FUSED_MAPPING = {
    "qkv_proj": ["q_proj", "k_proj", "v_proj"],
    "gate_up_proj": ["gate_proj", "up_proj"],
}

def _qwen3_next_ignored_layers(layer_idx: int, name: str) -> list:
    """
```

```
    模拟 Fp8Config.from_config 中的规范化：每个条目同时保留
    "model.<...>" 和 bare "<...>" 两种形式。
    """
    base = f"layers.{layer_idx}.linear_attn.{name}"
    return [base, f"model.{base}"]


class TestIsLayerSkipped(CustomTestCase):
    def test_qwen3_next_in_proj_ba_is_skipped(self):
        # 回归测试 for #23467：in_proj_ba 是 FP8 checkpoint 中的统一张量，
        # modules_to_not_convert 明确列出它，因此必须跳过 FP8 量化。
        # 否则在 tp=4 时 validate_block_quant_shapes 会因
        # output_partition_size=8 与 block_n=128 不兼容而报错。
        prefix = "model.layers.0.linear_attn.in_proj_ba"
        ignored = _qwen3_next_ignored_layers(0, "in_proj_ba")
        self.assertTrue(is_layer_skipped(prefix, ignored, QWEN3_NEXT_FUSED_MAPPING))

    def test_qwen3_next_in_proj_qkvz_is_skipped(self):
        prefix = "model.layers.5.linear_attn.in_proj_qkvz"
        ignored = _qwen3_next_ignored_layers(5, "in_proj_qkvz")
        self.assertTrue(is_layer_skipped(prefix, ignored, QWEN3_NEXT_FUSED_MAPPING))

    def test_mlp_gate_does_not_match_gate_up_proj(self):
        # 验证 #23467 的原始动机：modules_to_not_convert 中的 "mlp.gate"
        # 不能错误地跳过 "mlp.gate_up_proj"。
        ignored = ["mlp.gate"]
        self.assertFalse(
            is_layer_skipped("model.layers.0.mlp.gate_up_proj", ignored, {})
        )
        self.assertTrue(is_layer_skipped("model.layers.0.mlp.gate", ignored, {}))
```

评论区精华

唯一的 review 评论来自合并者 b8zhong，他批准了 PR 并询问："Do you mind adding a note somewhere in the code, since I think this is the difference of Q3N weight format vs Q3.5 weight format?" 这暗示了 Qwen3-Next 与 Qwen3.5 权重格式的差异应在代码中明确注释，以避免未来再次误操作。但该建议未在本次 PR 中实施。

- 代码注释建议：Qwen3-Next 与 Qwen3.5 权重格式差异 (other): 未在本次 PR 中实施，但建议在后续 PR 中补充注释。

风险与影响

- 风险：回归风险：修复移除两个条目后，若其他模型（如 Qwen3.5）的 FP8 配置确实依赖于这两个条目进行正确的分片映射，则可能导致类似加载失败。但 PR 说明指出，这些模型会通过自身的 packed_modules_mapping 声明融合关系，is_layer_skipped 优先使用该映射而非回退表，因此风险较低。测试覆盖：新增的回归测试覆盖了关键场景，但未测试 Qwen3.5 等模型是否仍能正常工作。建议在 CI 中补充运行相关模型的 FP8 加载测试。

- 影响：直接影响：修复了 Qwen3-Next FP8 在 B200 4-GPU 上无法启动的问题。影响范围：仅限于使用 _FALLBACK_FUSED_SHARDS 回退映射的量化配置；对于自带 packed_modules_mapping 的模型配置无影响。对用户：Qwen3-Next FP8 用户可正常加载模型。对系统：变更极小，仅删除两行代码，性能无影响。
- 风险标记：缺少注释说明权重格式差异，仅覆盖 Qwen3-Next 场景，未测试 Qwen3.5

关联脉络

- PR #23467 fix: dot-boundary match in is_layer_skipped for FP8 modules_to_not_convert: 本 PR 修复了 #23467 引入的回归问题。#23467 引入了 _FALLBACK_FUSED_SHARDS 并错误添加了 in_proj_ba 和 in_proj_qkvz。