

PR #35341 完整报告

sgl-project/sglang

[AMD][Fix] Qwen3.5: make empty-batch guard tuple-aware on fused AR+quant path

合并时间: 2026-08-28 16:11

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35341>

执行摘要

- 一句话: 修复 Qwen3.5 AMD 融合路径空批量守卫元组崩溃
- 推荐动作: 值得精读。该 PR 是一个典型的数据契约变更 (函数返回值从张量变为元组) 引起的下游兼容修复, 展示了如何通过最小改动适配上游变化。对于涉及此类融合 kernel 的开发, 有参考价值。

功能与动机

PR #24651 使 Qwen3.5 的 AMD 融合路径返回元组, 而 PR #34474 添加的空批量守卫直接读取 `hidden_states.shape[0]`, 导致守卫在 attention 解包元组之前抛出 `AttributeError`。问题在 CUDA graph 捕获时暴露, 表现为测试服务器以 `code -9` 退出。

实现拆解

1. 在 `qwen3_5.py` 的 GDN 层 `forward` 方法中, 将空批量守卫的条件从 `hidden_states.shape[0] > 0` 改为先解包: 若 `hidden_states` 是元组则取第一个元素, 再判断其 `shape[0]`, 同时传入 `linear_attn` 的仍是原始 `hidden_states`。
2. 在 `full-attention` 层的 `forward` 方法中做同样的修改, 因为该层接收的是二元组 (`fp8`, `scale`), 同样需要在守卫处解包。
3. 该实现参考了 `glm4_moe.py` 中的既有惯用法, 采用内联解包而非辅助函数, 保持 `diff` 最小化。
4. 未新增测试文件, 依赖已有的 `test_qwen35_fp8_ar_fusion_mi35x.py` 分区验证。

关键文件:

- `python/sglang/srt/models/qwen3_5.py` (模块 模型实现; 类别 `source`; 类型 `data-contract`; 符号 `Qwen3_5DecoderLayer.forward`, `Qwen3_5Attention.forward`) : 核心修复文件, 修改两处空批量守卫以兼容元组返回值。

关键符号: `Qwen3_5DecoderLayer.forward`, `Qwen3_5Attention.forward`

关键源码片段

`python/sglang/srt/models/qwen3_5.py`

核心修复文件, 修改两处空批量守卫以兼容元组返回值。

```
# python/sglang/srt/models/qwen3_5.py
```

```
# 空批量守卫需要兼容元组返回值:
# fused AR+quant 路径会返回 (fp8, scale) 或 (bf16, fp8, scale)
# 而 #34474 添加的守卫直接读取 hidden_states.shape[0], 导致崩溃
hs = hidden_states[0] if isinstance(hidden_states, tuple) else hidden_states
if not forward_batch.forward_mode.is_idle() and hs.shape[0] > 0:
    # 注意: 这里仍把原始 hidden_states (可能是元组) 传给 attention
    hidden_states = self.linear_attn(hidden_states, forward_batch)
```

评论区精华

Review 评论较少，无实质性讨论。主要审核意见来自合并者 HaiShaw 和审核者 mqhc2020 的 APPROVED，以及 amd-bot 的 CI 状态提醒。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。改动集中在两个守卫点，逻辑简单且不改变数据流。潜在风险：如果 `hidden_states` 元组内部第一个元素不是张量（如包含 `None`），仍会报错，但根据现有代码路径，元组内为张量。另外，该修复仅针对元组情况，若未来路径返回其他类型，仍需扩展。
- 影响：影响范围限于 Qwen3.5 模型在 AMD 环境（启用 fused AR+quant）下的推理路径。修复后，空批量守卫正确工作，避免 CUDA graph 捕获崩溃，提升该场景的稳定性。对 GPU（NV）等其他后端无影响。
- 风险标记：核心路径变更，依赖上游 PR，缺少测试覆盖

关联脉络

- PR #24651 AMD fused AR+quant returning tuples: 该 PR 引入了元组返回值，是本 PR 修复的根源。
- PR #34474 Add empty-batch guards on attention path: 该 PR 添加了空批量守卫，但未考虑元组情况，需本 PR 修复。
- PR #35250 Superseded partial fix for GDN guard: 本 PR supersedes 的旧修复，仅覆盖 GDN 守卫。