PR #28635 完整报告

sgl-project/sglang

[NPU] Add head_dim=256 to _can_use_tnd whitelist

合并时间：2026-06-18 17:55

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28635

执行摘要

- 一句话：为 Ascend NPU 添加 head_dim=256 的 TND 白名单支持
- 推荐动作：建议合入，变更简单且风险低。若团队后续深度使用该类模型，可考虑补充 TND 布局的单元测试或集成测试。

功能与动机

解决 Ascend NPU 上 head_dim=256 的模型被强制路由到较慢的 BSND per-sequence fallback 路径的问题。该 fallback 在 chunked prefill 结合 speculative decode 时存在单独缺陷。

实现拆解

1. 修改 AscendAttnBackend._can_use_tnd 静态方法，将 256 加入 d in (128, 192, 256) 的判断中。
2. 文件：python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py，仅一行变更。
3. 无测试、配置或部署配套修改。

关键文件：

- python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py（模块 NPU 注意力后端；类别 source；类型 core-logic；符号 _can_use_tnd）：核心变更文件，在 _can_use_tnd 方法中加入 head_dim=256 支持。

关键符号：_can_use_tnd

关键源码片段

python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py

核心变更文件，在 _can_use_tnd 方法中加入 head_dim=256 支持。

```
# python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py

@staticmethod
def _can_use_tnd(layer: RadixAttention) -> bool:
    """Check if TND layout is supported."""
```

```
    d = layer.qk_head_dim
    v = layer.v_head_dim
    # 允许 head_dim=256 使用 TND 布局（CANN 官方仅承诺 128/192，但实测可用）
    return (d == v and d in (128, 192, 256)) or (d == 192 and v == 128)
```

评论区精华

无实质性技术讨论。只有 Gemini Code Assist 机器人自动评论确认无反馈，以及 CI 机器人自动批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅放宽了 TND 布局使用的条件，且已在两不同架构（Qwen3.5 和 Gemma4）上验证过。但未添加新测试，未来 head_dim=256 模型若出现精度问题，可能被当前变更掩盖。
- 影响：影响范围小，仅涉及 Ascend NPU 上 head_dim=256 的模型。从缓慢的 BSND fallback 切换到 TND fast path，可提升 prefill 性能并规避已知 bug。
- 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR