PR #26996 完整报告

sgl-project/sglang

Fix dp-attention token alignment in the dumper comparator e2e test

合并时间：2026-06-02 15:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26996

执行摘要

- 一句话：修复 dp-attention e2e comparator 测试 token 对齐失败
- 推荐动作：值得快速合入。该 PR 体现了对 dp-attention 跨 rank 数据布局的深入理解，修复思路清晰。对于涉及多 GPU 分布式推理的开发者，可借此 PR 了解 comparator 的 token aligner 机制。

功能与动机

test_dp_attention 用例（nightly-4-gpu）持续失败：comparator 报告每个 tensor 的 shape_mismatch（passed=0, failed=21）。根本原因是 dp-attention 在 dump 点之前跨 DP rank 聚集 token，单请求下空 DP rank 贡献一个 padding token，导致非空 rank 的 gathered buffer 尺寸为 [6,] 而 baseline 为 [5,]，shape 不匹配。

实现拆解

1. 根因分析：在 dp-attention 模式下（attn_tp_size=1, attn_dp_size=2），tokens 在 dump 点前跨 rank 聚集；单请求时一个 DP rank 为空，其贡献的 padding 导致 shape 偏差。
2. 修复方式：在 test_dp_attention 方法的 extra_comparator_args 中添加 --token-aligner concat_steps。该 aligner 从 dumped seq-lens 重构真实每步 token 序列并裁剪 padding，使 target 与未补零的 baseline 对齐。
3. 配套说明：在 docstring 中新增段落解释 padding 产生的机制及 aligner 的工作原理，便于后续维护。
4. 验证：在 4-GPU H200 上针对 Qwen/Qwen3-30B-A3B 运行两个 e2e 用例均通过；在官方 CI 4-gpu-h100 runner 上也验证通过。

关键文件：

- test/registered/debug_utils/test_engine_dumper_comparator_e2e.py（模块 测试框架；类别 test；类型 test-coverage；符号 TestSourcePatcherE2ESGLang.test_dp_attention）：唯一变更文件：在 test_dp_attention 用例的 comparator 参数中添加 --token-aligner concat_steps，并更新 docstring 说明 padding 原因。

关键符号：TestSourcePatcherE2ESGLang.test_dp_attention

关键源码片段

test/registered/debug_utils/test_engine_dumper_comparator_e2e.py

唯一变更文件：在 test_dp_attention 用例的 comparator 参数中添加 --token-aligner concat_steps，并更新 docstring 说明 padding 原因。

```
def test_dp_attention(self, tmp_path: Path) -> None:
    """TP=2 baseline vs TP=2+DP=2+dp-attention target.

    # ... 原有注释 ...

    The ``concat_steps`` token aligner is required because dp-attention
    gathers tokens across DP ranks before the dump point, so the
    non-empty rank's buffer holds the real tokens plus padding tokens
    contributed by the empty DP rank (with a single request one DP
    rank is always empty). The aligner reconstructs the real per-step
    token sequence from the dumped seq-lens, trimming that padding so
    the target lines up with the un-padded TP baseline.
    """
    _run_e2e_scenario(
        tmp_path=tmp_path,
        target_tp=BASELINE_TP,
        extra_target_server_args=["--dp", "2", "--enable-dp-attention"],
        target_patch_config_yaml=PATCH_CONFIG_DP_ATTENTION_YAML,
        extra_comparator_args=[
            "--token-aligner", # 新增：使用 concat_steps 重构真实 token 序列
            "concat_steps",
            "--end-step",
            "0",
            "--allow-failed-pattern",
            "mlp_output",
        ],
    )
```

评论区精华

无实质性 review 讨论，alisonshao 直接批准。PR 作者在 issue 评论中详细记录了调查过程和验证结果。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅对测试文件新增一个 aligner 参数和 docstring，不影响任何生产代码逻辑。concat_steps aligner 为已有功能，专为 BS=1 场景设计，与 dp-attention 用例的单请求前提一致。
- 影响：仅影响 test_engine_dumper_comparator_e2e.py 中 test_dp_attention 用例的比对逻辑；修复后该用例可从持续失败转为通过，消除 nightly 4-GPU 测试噪声。不影响其他测试或生产系统。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR