PR #24241 完整报告

sgl-project/sglang

[bugfix] Support MIXED forward mode in TBO splitter for DP attention

合并时间：2026-05-02 07:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24241

执行摘要

- 一句话：修复 DP attention 与 mixed chunk 组合时的崩溃
- 推荐动作：值得合入，修复严重崩溃 bug，改动极小且带有回归测试。可关注后续是否将 --enable-two-batch-overlap 与 mixed chunk 的支持补全。

功能与动机

--enable-dp-attention 与 --enable-mixed-chunk 组合会确定性崩溃，因为 compute_split_seq_index 和 compute_split_token_index 未处理 ForwardMode.MIXED。PR body 指出该组合之前被静默破坏，server_args.py 中没有任何验证阻止它，但一旦调度器进入 MIXED 模式就会崩溃。

实现拆解

1. 修改 split 逻辑：在 python/sglang/srt/batch_overlap/two_batch_overlap.py 的 compute_split_seq_index（第 84 行）和 compute_split_token_index（第 273 行）中，将条件从 forward_mode == ForwardMode.EXTEND 扩展为 forward_mode == ForwardMode.EXTEND or forward_mode == ForwardMode.MIXED，使 MIXED 模式复用 EXTEND 的分割逻辑。
2. 原理说明：mix_with_running 操作后，running decode 请求被追加到 extend_lens 作为长度为 1 的条目，因此 _split_extend_seqs 和累积和分割逻辑可直接复用。
3. 添加回归测试：在 test/registered/distributed/test_dp_attention.py 中新增 TestDPAttentionMixedChunk 测试类，继承 CustomTestCase 和 GSM8KMixin，启动服务器时传入 --enable-dp-attention --dp 2 --enable-mixed-chunk --chunked-prefill-size 256 参数，并通过 GSM8K 准确率阈值 0.6 确保推理正确性。

关键文件：

- python/sglang/srt/batch_overlap/two_batch_overlap.py（模块 调度器；类别 source；类型 core-logic；符号 compute_split_seq_index, compute_split_token_index）：核心 bugfix 文件，修改 compute_split_seq_index 和 compute_split_token_index 以处理 MIXED forward mode
- test/registered/distributed/test_dp_attention.py（模块 DP 注意力；类别 test；类型 test-coverage；符号 TestDPAttentionMixedChunk, setUpClass, tearDownClass）：新增回归测试类 TestDPAttentionMixedChunk，验证 DP attention + mixed chunk 组合的

正确性

关键符号：compute_split_seq_index, compute_split_token_index

关键源码片段

python/sglang/srt/batch_overlap/two_batch_overlap.py

核心 bugfix 文件，修改 compute_split_seq_index 和 compute_split_token_index 以处理 MIXED forward mode

```
def compute_split_seq_index(
    forward_mode: ForwardMode,
    num_tokens: int,
    extend_lens: Optional[Sequence[int]],
    token_num_per_seq: Optional[int],
) -> Optional[int]:
    # 关键变更：将 MIXED 模式视为 EXTEND，因为 mix_with_running 后
    # running decode 请求被追加为长度 1 的 extend_lens
    if forward_mode == ForwardMode.EXTEND or forward_mode == ForwardMode.MIXED:
        assert extend_lens is not None
        return _split_extend_seqs(extend_lens)
    elif forward_mode.is_target_verify() or forward_mode.is_decode():
        assert token_num_per_seq is not None
        return (num_tokens // token_num_per_seq) // 2
    elif forward_mode.is_idle() or forward_mode.is_prebuilt():
        assert num_tokens == 0
        return 0
    else:
        raise NotImplementedError()


def compute_split_token_index(
    split_seq_index: int,
    forward_mode: "ForwardMode",
    extend_seq_lens: Optional[Sequence[int]],
    token_num_per_seq: Optional[int],
) -> int:
    # 同样处理 MIXED 模式
    if forward_mode == ForwardMode.EXTEND or forward_mode == ForwardMode.MIXED:
        assert extend_seq_lens is not None
        if _is_two_chunk_split_enabled(extend_seq_lens):
            return sum(extend_seq_lens) // 2
        return sum(extend_seq_lens[:split_seq_index])
    elif forward_mode.is_target_verify() or forward_mode.is_decode():
        assert token_num_per_seq is not None
        return split_seq_index * token_num_per_seq
    elif forward_mode.is_idle():
        assert split_seq_index == 0
        return 0
```

```
        else:
            raise NotImplementedError
```

test/registered/distributed/test_dp_attention.py

新增回归测试类 TestDPAttentionMixedChunk，验证 DP attention + mixed chunk 组合的正确性

```
class TestDPAttentionMixedChunk(
    CustomTestCase,
    GSM8KMixin,
):
    # 设置 GSM8K 准确率阈值为 0.6，用于验证推理正确性
    gsm8k_accuracy_thres = 0.6

    @classmethod
    def setUpClass(cls):
        cls.model = DEFAULT_MLA_MODEL_NAME_FOR_TEST
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=[
                "--trust-remote-code",
                "--tp", "2",
                "--enable-dp-attention",
                "--dp", "2",
                "--enable-mixed-chunk", # 之前会崩溃的选项
                "--chunked-prefill-size", "256", # 触发 chunked prefill
            ],
        )

    @classmethod
    def tearDownClass(cls):
        kill_process_tree(cls.process.pid)
```

评论区精华

gemini-code-assist[bot] 建议将 forward_mode == ForwardMode.EXTEND or forward_mode == ForwardMode.MIXED 改为 forward_mode in (ForwardMode.EXTEND, ForwardMode.MIXED) 以提高可读性。该建议未被采纳，但属于风格优化，不影响功能正确性。

- 使用 in 运算符简化枚举比较 (style): 未采纳，但属于风格优化，不影响功能。

风险与影响

- 风险：风险较低。变更仅在两处条件判断中添加 MIXED 枚举值匹配，逻辑路径与 EXTEND 一致，且通过 assert extend_lens is not None 保证前置条件。未改动的 OperationsStrategy.init_new_tbo 对 MIXED 仍会抛出 NotImplementedError，但该路径

仅在 --enable-two-batch-overlap 生效时到达，而该组合仍被标记为不支持，因此无回归风险。

- 影响：修复了 --enable-dp-attention 与 --enable-mixed-chunk 的组合崩溃 bug，使 DP attention 用户可以使用 mixed chunk 功能，提升吞吐。影响范围限定于使用这两个选项的 DP attention 场景，且不涉及 Two Batch Overlap 路径。
- 风险标记：无显著风险

关联脉络

- 暂无明显关联 PR