PR #33953 完整报告

sgl-project/sglang

[Diffusion] fix: scope the masked-path replicated guard to SP runs

合并时间：2026-08-07 15:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33953

执行摘要

- 一句话：SP 守卫限定，修复单秩 masked 路径 CI 失败
- 推荐动作：值得精读。这是一个小而关键的回归修复，展示了“守卫条件必须与实际运行模式对齐”的工程原则。阅读 layer.py 的守卫逻辑变化和新增测试，可学习如何在分布式与单秩之间精确收窄保护范围。建议快速合并，并关注后续 SP 场景的回归测试。

功能与动机

PR body 指出 #33923 合入的 masked-path guard 会拒绝单秩调用，但单秩下 mask 已描述完整序列，replicated 计数是无操作；qwen 整个 1-GPU masked 家族（t2i/edit/2509/2511/layered）因此在 component-accuracy 中抛出 NotImplementedError，且所有打开 PR 的 CI 快速失败。需要将守卫精确限定到 SP 运行，以恢复 CI 并允许合法单秩调用。

实现拆解

1. 分析守卫条件：在 python/sglang/multimodal_gen/runtime/layers/attention/layer.py 的 USPAttention.forward 中，原有条件只要 attn_mask 存在且任一 replicated 计数非零即抛 NotImplementedError，未考虑 SP 世界大小，导致单秩误拒绝。
2. 收窄守卫范围：在条件中追加 and not effective_skip_sp and get_sequence_parallel_world_size() > 1，使守卫仅在 SP 实际运行且未跳过 SP 时生效；同步更新错误信息，明确“在序列并行下”的限制。这样单秩（world_size == 1）和明确跳过 SP 的调用保持合法。
3. 新增回归测试：在 python/sglang/multimodal_gen/test/unit/test_usp_attention_replicated_prefix.py 中添加 test_single_rank_masked_call_keeps_replicated_args，mock get_sequence_parallel_world_size 返回 1，验证带 num_replicated_prefix=2 的 masked 调用不再抛错且输出形状正确。测试同时补充了 AttentionBackendEnum 导入及构造 SDPA 所需属性（allow_cudnn_sdp、softmax_scale、backend 等）。
4. 验证 CI：作者在 issue 评论中报告 component-accuracy 套件已恢复绿色（29m57s 通过），并确认其他平台仅剩 AMD/NPU 环境抖动，成功解除多个 PR 的 CI 阻塞。

关键文件：

- python/sglang/multimodal_gen/runtime/layers/attention/layer.py（模块 注意力层；类别 source；类型 core-logic；符号 USPAttention.forward）：核心逻辑修改：将

masked-path replicated 守卫限定到 SP 运行，避免单秩误拒绝。

- python/sglang/multimodal_gen/test/unit/test_usp_attention_replicated_prefix.py（模块 USP 注意力；类别 test；类型 test-coverage；符号 test_single_rank_masked_call_keeps_replicated_args）：新增单秩 masked 调用回归测试，覆盖修复路径，防止未来误拒绝。

关键符号：USPAttention.forward, test_single_rank_masked_call_keeps_replicated_args

关键源码片段

python/sglang/multimodal_gen/runtime/layers/attention/layer.py

核心逻辑修改：将 masked-path replicated 守卫限定到 SP 运行，避免单秩误拒绝。

```
# 在 USPAttention.forward 中，masked 路径的 replicated 守卫：
if attn_mask is not None or meta_only_pad:
    # 仅当实际运行序列并行（SP）且未跳过 SP 时，masked 路径才会按 rank
    # 通过 all-to-all 分片每一行；此时 replicated prefix/suffix 会被复制到
    # 各 rank，导致输出静默损坏，因此必须显式报错。
    # 单 rank 场景下 mask 已描述完整序列，replicated 计数是无效的 no-op，
    # 调用必须保持合法。
    if (
        (num_replicated_prefix or num_replicated_suffix or num_replicated_kv_prefix)
        and not effective_skip_sp
        and get_sequence_parallel_world_size() > 1
    ):
        raise NotImplementedError(
            "USPAttention's masked path does not support replicated "
            "prefix/suffix tokens under sequence parallelism; drop "
            "attn_mask/attn_mask_meta or the replicated segment."
        )
```

python/sglang/multimodal_gen/test/unit/test_usp_attention_replicated_prefix.py

新增单秩 masked 调用回归测试，覆盖修复路径，防止未来误拒绝。

```
class TestUSPAttentionMaskedReplicatedGuard(unittest.TestCase):
    # ... 原有 test_masked_path_rejects_replicated_tokens 保留 ...

    def test_single_rank_masked_call_keeps_replicated_args(self):
        # 无 SP 时 mask 描述完整序列，replicated 计数无意义，必须放行
        obj = USPAttention.__new__(USPAttention)
        obj.attn_impl = _CaptureAttn()
        obj.skip_sequence_parallel = False
        obj.sp_attention_mode = "ulysses"
        obj.sp_attention_mode_is_auto = False
        obj.allow_cudnn_sdp = False
        obj.softmax_scale = 0.5
        obj.backend = AttentionBackendEnum.TORCH_SDPA
        obj.causal = False
        obj.dropout_p = 0.0
```

```
        q = torch.randn(1, 6, 2, 4)
        mask = torch.ones(1, 6, dtype=torch.bool)
        with (
            patch(
                f"{_LAYER}.get_forward_context",
                return_value=MagicMock(attn_metadata=None),
            ),
            patch(f"{_LAYER}.get_sequence_parallel_world_size", return_value=1),
        ):
            out = obj.forward(q, q, q, attn_mask=mask, num_replicated_prefix=2)
        self.assertEqual(out.shape, q.shape)
```

评论区精华

PR 没有独立 review 评论，但作者在 issue 评论中补充了关键结论："CI settled: component-accuracy is green again on this branch ([29m57s pass]...)—the exact suite that has been red on main since #33923 merged, and zero NV-side real failures overall（remaining reds are the AMD/NPU platform flakes every PR currently shows）. This unblocks every open PR's fast-fail cascade (e.g. #33928's 2-gpu (0)), so a quick merge would be appreciated." 讨论核心：修复了一个由 #33923 引入的 CI 回归，且验证了修复没有引入新的 NV-side 失败。

- 单秩 masked 调用被非法拒绝导致 CI 失败 (correctness): 将守卫限定到 SP 运行（world size > 1 且未 skip SP），单秩场景合法；新增回归测试验证。

风险与影响

- 风险：风险很低，但需关注两点： 1) 守卫现在只在 get_sequence_parallel_world_size() > 1 且 effective_skip_sp 为假时触发，若未来 SP 相关逻辑变化（如 world size 在 forward 中途变化），守卫可能失效； 2) 错误信息文本变更可能影响依赖该字符串的测试或外部断言，但仓库内未见其他引用。文件改动集中在 2 个文件，回归面小。
  - 影响：影响范围：解除 main 分支上所有打开 PR 的 CI 快速失败（如 #33928），恢复 qwen 1-GPU masked 系列的 component-accuracy 通过；对用户无直接运行时影响（修复对象是错误拒绝）。团队可恢复正常 PR 合入流程。长期看，守卫的 SP 限定使单秩 masked 路径保持可用，双秩 SP 仍受保护。
  - 风险标记：CI 阻塞解除，SP 世界大小依赖

关联脉络

- PR #33923 [Diffusion] Route zimage and hunyuanvideo attention through USPAttention: 该 PR 引入了被本 PR 修复的 masked-path replicated 守卫，是回归的直接来源。