PR #44130 完整报告

vllm-project/vllm

[Bugfix] Fix `sequence_parallel_chunk_impl` custom op aliasing its input

合并时间：2026-06-06 07:56

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44130

执行摘要

- 一句话：修复 custom op 返回视图导致启动崩溃
- 推荐动作：该 PR 改动简单但关键，建议所有使用 sequence-parallel MoE 的部署及时合入。可关注后续是否需要在测试中加入针对 custom op 输出别名的静态检查，防止同类问题。

功能与动机

MoE 模型启用 sequence-parallel MoE（ --enable-expert-parallel 等配置）时，引擎启动报错：RuntimeError: The output of this custom operator must not alias any inputs。该问题阻断了 Qwen3.5/Qwen3-Next 等模型的正常使用。

实现拆解

1. 在 vllm/model_executor/models/utils.py 的 sequence_parallel_chunk_impl 函数中，将返回值拆分为两步：先通过 torch.narrow 切片，然后判断 y is x（即未 padding 的情况），若为真则调用 clone()，否则直接返回切片。
2. 修改仅影响自定义算子 vllm::sequence_parallel_chunk_impl 的实现，对外接口不变。
3. 未新增测试文件，但作者提供了详细的复现命令及其在 4x B200 上的成功验证结果。

关键文件：

- vllm/model_executor/models/utils.py（模块 模型执行；类别 source；类型 data-contract；符号 sequence_parallel_chunk_impl）：包含 sequence_parallel_chunk_impl 自定义算子的实现，是 bug 根因和修复位置。

关键符号：sequence_parallel_chunk_impl

关键源码片段

vllm/model_executor/models/utils.py

包含 sequence_parallel_chunk_impl 自定义算子的实现，是 bug 根因和修复位置。

```
# 关键函数： sequence_parallel_chunk_impl
# 该函数被注册为 torch custom op，AOT-autograd 要求输出不能别名输入。
# 当序列长度恰好能被 tp_size 整除时，无需 padding，y = x
# 此时 torch.narrow 返回的是 x 的视图，因此需要 clone 确保独立性。
def sequence_parallel_chunk_impl(x: torch.Tensor) -> torch.Tensor:
```

```
tp_size = get_tensor_model_parallel_world_size()
tp_rank = get_tensor_model_parallel_rank()
seq_len = x.size(0)
remainder = seq_len % tp_size
if remainder != 0:
    pad_len = tp_size - remainder
    y = nn.functional.pad(x, (0, 0, 0, pad_len)) # padding 分支，y 是新张量
else:
    y = x # no-pad 分支，y 是输入别名
chunk = y.shape[0] // tp_size
start = tp_rank * chunk
out = torch.narrow(y, 0, start, chunk)
# 检查是否别名，若是则 clone，满足 custom op 契约
return out.clone() if y is x else out
```

评论区精华

issue 评论中 ZJY0516 表示本地无法复现该问题，仅看到 warning 而非报错，说明问题可能依赖于特定 PyTorch 版本或 compile 配置。PR 本身无 review 评论，直接获得 approval。

- 暂无高价值评论线程

风险与影响

- 风险：风险低。修改仅增加一次 clone 操作，在 no-pad 路径下引入额外内存拷贝，但该路径本身不含 pad 的显存开销，整体影响微弱。未引入新的配置或依赖；已通过实际用例验证。
- 影响：直接修复了使用 sequence-parallel MoE 的模型（如 Qwen3.5-397B-A17B-NVFP4）的引擎启动崩溃问题。影响范围限于 tensor_parallel_size > 1, data_parallel_size > 1, allgather_reducescatter 后端场景。用户无需修改代码或配置即可受益。
- 风险标记：核心路径变更

关联脉络

- PR #44561 [DSV4] Move more ops out of eager breakpoint: 同为 MoE 模型 sequence parallel 优化相关，涉及相似模块