PR #47050 完整报告

vllm-project/vllm

[BugFix] Gate MRV2 mixed sparse-MLA warmup on `max_num_seqs` > 1

合并时间：2026-06-30 23:31

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/47050

执行摘要

- 一句话：适配 MRV2 warmup 在单序列场景的跳过逻辑
- 推荐动作：值得快速合入。修复清晰，逻辑简单，建议在合入后观察 CI 测试稳定性。

功能与动机

由 @ZeldaHuang 发现，当 max_num_reqs <= 1 时，混合 prefill+decode warmup 会尝试构造一个 prefill 和一个 decode 请求，但受限于 max_num_reqs 限制导致失败。此修复添加了前置检查，避免无效执行。

实现拆解

1. 核心门控加固：在 vllm/v1/worker/gpu/warmup.py 的 run_mixed_prefill_decode_warmup 函数中，将早期返回条件从 is_pooling_model or num_tokens < 3 扩展为 is_pooling_model or max_num_reqs < 2 or num_tokens < 3。
2. 调用方同步：在 vllm/model_executor/warmup/flashinfer_sparse_mla_warmup.py 中，对两处 V2 warmup 调用（_run_flashinfer_sparse_mla_decode_autotune 和 deepseek_v4_sparse_mla_attention_warmup）增加 runner.max_num_reqs >= 2 的条件检查，避免逻辑上绕开门控。
3. 配套测试：新增 tests/v1/worker/test_mixed_warmup_gate.py，使用 SimpleNamespace 模拟 runner，验证 max_num_reqs 为 0 或 1 时 warmup 返回 False 且 worker 回调不会被调用。

关键文件：

- vllm/v1/worker/gpu/warmup.py（模块 预热逻辑；类别 source；类型 core-logic）：核心门控函数 run_mixed_prefill_decode_warmup 的早期返回条件增加 max_num_reqs < 2 检查。
- vllm/model_executor/warmup/flashinfer_sparse_mla_warmup.py（模块 模型预热；类别 source；类型 data-contract）：两处调用 V2 warmup 的位置添加 runner.max_num_reqs >= 2 条件，确保调用时机正确。
- tests/v1/worker/test_mixed_warmup_gate.py（模块 测试；类别 test；类型 test-coverage；符号 _fail, test_mixed_warmup_skipped_for_single_seq）：新增测试文件，验证 max_num_reqs <= 1 时 warmup 被跳过且不调用 worker 回调。

关键符号：run_mixed_prefill_decode_warmup, _run_flashinfer_sparse_mla_decode_autotune, deepseek_v4_sparse_mla_attention_warmup

关键源码片段

vllm/v1/worker/gpu/warmup.py

核心门控函数 run_mixed_prefill_decode_warmup 的早期返回条件增加 max_num_reqs < 2 检查。

```
def run_mixed_prefill_decode_warmup(
    model_runner: GPUModelRunner,
    worker_execute_model: Callable[[SchedulerOutput], Any],
    worker_sample_tokens: Callable[[GrammarOutput | None], Any],
    num_tokens: int,
    *,
    mixed_step_context: AbstractContextManager[object] | None = None,
    req_id_prefix: str = "_v2_mixed_warmup",
) -> bool:
    # 混合 prefill+decode 步骤需要至少 2 个请求：
    # 一个 prefill 和一个 decode；当 max_num_reqs < 2 时直接跳过
    if model_runner.is_pooling_model or model_runner.max_num_reqs < 2 or num_tokens < 3:
        return False
    # 后续构造请求并执行模型
    ...
```

vllm/model_executor/warmup/flashinfer_sparse_mla_warmup.py

两处调用 V2 warmup 的位置添加 runner.max_num_reqs >= 2 条件，确保调用时机正确。

```
# 在 _run_flashinfer_sparse_mla_decode_autotune 中：
if _uses_v2_model_runner(runner) and runner.max_num_reqs >= 2:
    v2_runner = cast("V2GPUModelRunner", runner)
    warmup_executed = run_mixed_prefill_decode_warmup(...)
else:
    runner._dummy_run(...)

# 在 deepseek_v4_sparse_mla_attention_warmup 中同样：
if _uses_v2_model_runner(runner) and runner.max_num_reqs >= 2:
    v2_runner = cast("V2GPUModelRunner", runner)
    run_mixed_prefill_decode_warmup(...)
else:
    runner._dummy_run(...)
```

评论区精华

PR 无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅添加一个前置条件检查，所有改动点均在 warmup 路径，不影响推理核心流程；测试覆盖边界场景。
- 影响：影响范围小。仅影响 V2 model runner 下 max_num_seqs = 1 的场景（如小批量在线服务或调节测试）。修复后这些场景不再触发无效 warmup，避免潜在错误日志或崩溃。
- 风险标记：低风险

关联脉络

- 暂无明显关联 PR