PR #31332 完整报告

sgl-project/sglang

[CI] Fix TRTLLM MHA graph metadata test fixture

合并时间：2026-07-16 03:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/31332

执行摘要

- 一句话：修复 TRTLLM MHA 图元数据测试 fixture
- 推荐动作：该 PR 为小型测试修复，值得查看以了解如何在图捕获测试中模拟 replay 状态变化。

功能与动机

PR#31013 修改了 num_tokens_per_req 的推导逻辑，导致测试中的 fixture 不再符合预期行为。测试原本设置 num_tokens_per_req=0，但实际 draft extend 阶段应有 4 个 token，因此需要更新 fixture 并添加额外的断言。

实现拆解

1. 在 test/registered/attention/test_trtllm_mha_graph_metadata.py 的 test_draft_extend_in_graph_uses_captured_static_q_stride 测试中，将 spec_info.num_tokens_per_req 的初始值从 0 改为 4，以匹配实际的 draft token 数。
2. 在 init_forward_metadata_out_graph 调用后、init_forward_metadata_in_graph 调用前，将 fb.spec_info.num_tokens_per_req 重置为 0，以模拟 replay 时该字段可能被修改的情况。
3. 添加注释说明 in-graph 体必须使用捕获的静态 stride，而非 replay 时的状态。

关键文件：

- test/registered/attention/test_trtllm_mha_graph_metadata.py（模块 测试；类别 test；类型 test-coverage）：测试修复的核心文件，修改了 fixture 的初始值并增加了状态重置逻辑。

关键符号：test_draft_extend_in_graph_uses_captured_static_q_stride

关键源码片段

test/registered/attention/test_trtllm_mha_graph_metadata.py

测试修复的核心文件，修改了 fixture 的初始值并增加了状态重置逻辑。

```
# test/registered/attention/test_trtllm_mha_graph_metadata.py

def test_draft_extend_in_graph_uses_captured_static_q_stride(monkeypatch):
```

```
calls = []

def fake_update(**kwargs):
    calls.append(kwargs)

class ExplodingAcceptTokens:
    def __getitem__(self, key):
        raise AssertionError("in-graph metadata must not inspect accept tokens")

monkeypatch.setattr(
    trtllm_mha_backend, "update_trtllm_mha_graph_metadata", fake_update
)
backend = _make_backend_for_hook_test(speculative_num_draft_tokens=4)
fb = SimpleNamespace(
    batch_size=2,
    req_pool_indices=torch.arange(2, dtype=torch.int64),
    seq_lens=torch.ones(2, dtype=torch.int32),
    forward_mode=ForwardMode.DRAFT_EXTEND_V2,
    spec_info=SimpleNamespace(
        num_tokens_per_req=4, # 改为与实际 draft token 数匹配
        num_accept_tokens=ExplodingAcceptTokens(),
    ),
    positions=torch.arange(8, dtype=torch.int64),
    out_cache_loc=torch.arange(8, dtype=torch.int64),
)

backend.init_forward_metadata_out_graph(fb, in_capture=True)
# The in-graph body must use the captured static stride, not replay-time state.
fb.spec_info.num_tokens_per_req = 0 # 模拟 replay 时该字段可能被修改
backend.init_forward_metadata_in_graph(fb)

assert len(calls) == 1
assert calls[0]["q_mode"] == Q_MODE_STRIDED
assert calls[0]["q_stride"] == 4
```

评论区精华

无 review 评论，仅一位贡献者提交变更，后由 b8zhong 触发 rerun 测试，测试通过。

- 暂无高价值评论线程

风险与影响

- 风险：变更仅涉及测试 fixture，无生产代码改动，风险极低。但需确保 PR#31013 的变更不会在其他测试中引入类似问题。
- 影响：直接影响：修复一个测试用例，使其通过 CI。间接影响：确保 TRTLLM MHA 图元数据在 speculative decoding 场景下的正确性验证。
- 风险标记：暂无

关联脉络

- PR #31013 [Spec] Single-source num_tokens_per_req derivation and access: 该 PR 的变更导致了测试 fixture 不匹配，本 PR 修复了因此引入的测试失败。