PR #26669 完整报告

sgl-project/sglang

test: add trtllm_mha EAGLE-draft CG runner coverage (chain)

合并时间：2026-05-29 14:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26669

执行摘要

- 一句话：为 TRTLLM MHA 添加 EAGLE-draft CG 测试覆盖
- 推荐动作：本次 PR 值得关注其测试方法论：如何通过注入历史 bug 来验证新测试的有效性。建议未来类似修复（尤其涉及 CG capture/replay 路径）都配套此类测试，并利用 bug 注入确保测试能真正捕获回归。

功能与动机

PR #26521 和 #26655 修复了 trtllm_mha_backend.py draft-decode CG 路径中的 bug，但由于没有测试覆盖该路径，这些 bug 在 CI 中未被检测到。PR body 指出："Both fixes shipped under green unittests because no test exercised that code path"，因此需要新增测试来驱动这些代码路径，防止未来回归。

实现拆解

1. 新增导入和测试数据：在 test/registered/attention/unittests/dense/test_trtllm_mha.py 中导入 run_dense_eagle_draft_cuda_graph_runner_case，并定义 EAGLE_DRAFT_RUNNER_CASES 元组，包含一个链式（topk=1）测试用例，使用 num_heads=4、num_kv_heads=4、page_size=16、prefix_lens=(4, 7) 以及 topk=1、num_draft_tokens=3。
2. 新增测试方法：test_runner_mode_eagle_draft_cuda_graph_runner_cases 方法遍历上述用例，调用 run_dense_eagle_draft_cuda_graph_runner_case，传递 topk 和 speculative_num_draft_tokens，直接覆盖 trtllm_mha_backend.py 中第 320-340 行的 capture 路径和第 460-476 行的 replay 路径。
3. 更新文档：修改 test/registered/attention/unittests/KNOWN_FAILURES.md，更新 H200 参考运行的测试统计（172 passed, 23 skipped, 536 subtests passed），并更新日期。同时修改 dense/README.md，将 trtllm_mha 行中 EAGLE-draft runner 列从 deferred: requires chain-only graph capture wiring 改为 ✓ chain (topk=1)。
4. 验证：全量 attention-unittest 套件 172 passed, 23 skipped, 536 subtests passed，无回归；通过暂逆 PR #26521 确认新测试失败，恢复后通过，验证了测试的有效性。

关键文件：

- test/registered/attention/unittests/dense/test_trtllm_mha.py（模块 测试；类别 test；类型 test-coverage；符号 test_runner_mode_eagle_draft_cuda_graph_runner_cases）：

核心变更文件：新增 EAGLE draft CUDA Graph runner 测试覆盖，包括导入、测试用例定义和测试方法实现，直接驱动之前未被测试的 CG capture/replay 路径。

- test/registered/attention/unittests/KNOWN_FAILURES.md（模块 文档；类别 docs；类型 documentation）：文档更新：反映新增测试后的 H200 参考运行统计和日期。
- test/registered/attention/unittests/dense/README.md（模块 文档；类别 docs；类型 documentation）：文档更新：将 trtllm_mha 的 EAGLE-draft runner 单元格从 deferred 改为 ✓ chain (topk=1)，反映测试覆盖状态。

关键符号：test_runner_mode_eagle_draft_cuda_graph_runner_cases

关键源码片段

test/registered/attention/unittests/dense/test_trtllm_mha.py

核心变更文件：新增 EAGLE draft CUDA Graph runner 测试覆盖，包括导入、测试用例定义和测试方法实现，直接驱动之前未被测试的 CG capture/replay 路径。

```
# test/registered/attention/unittests/dense/test_trtllm_mha.py
# 新增导入：EAGLE draft runner 测试辅助函数
from sglang.test.kits.attention_unittest.runner_modes.speculative_draft_runner import (
    run_dense_eagle_draft_cuda_graph_runner_case,
)

# ...

# EAGLE draft CG runner — chain only (topk=1). trtllm_mha is constrained
# to topk=1 via `trtllm_mha_backend.py:459,492` so tree-mode tests don't
# apply. This test exercises the draft-decode CG capture/replay path
# (`init_forward_metadata_capture_cuda_graph` line 320 and
# `init_forward_metadata_replay_cuda_graph` line 460) — the same path
# patched by PR #26521 (capture-time NaN fix) and PR #26655 (replay-time
# slice rebind).
EAGLE_DRAFT_RUNNER_CASES = (
    (
        DenseAttentionCase(
            name="runner_eagle_draft_decode_trtllm_mha_cuda_graph_chain",
            backend="trtllm_mha",
            forward_mode=ForwardMode.DECODE,
            num_heads=4,
            num_kv_heads=4,
            page_size=16,
            prefix_lens=(4, 7),
        ),
        1, # topk — chain only; tree mode (topk >= 2) not supported
        3, # num_draft_tokens
    ),
)

def test_runner_mode_eagle_draft_cuda_graph_runner_cases(self):
```

```
    for case, topk, num_draft_tokens in self.EAGLE_DRAFT_RUNNER_CASES:
        with self.subTest(case=case.name, backend=case.backend, topk=topk):
            run_dense_eagle_draft_cuda_graph_runner_case(
                self,
                case,
                topk=topk,
                speculative_num_draft_tokens=num_draft_tokens,
                head_dim=self.HEAD_DIM,
                hidden_size=self.HIDDEN_SIZE,
            )
```

评论区精华

无实质性 review 讨论。chatgpt-codex-connector[bot] 的自动评论提及了 speculative_cuda_graph_runner.py 中 prod_fill padding 相关的一个 P2 级别建议，与本次新增测试用例无直接冲突。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。本次变更仅涉及测试文件（新增 38 行，删除 0 行）和文档更新（共 6 行），不修改任何生产代码。新增测试通过 bug 注入验证确认有效，不会引起回归。唯一的轻微风险是测试耗时增加约 0.26%（原 536 子测试 +1），但已被 CI 接受。
- 影响：影响范围仅限于测试套件和文档。对最终用户无影响，对系统运行无影响。对开发团队的意义在于：为 trtllm_mha 的 EAGLE-draft CG 路径提供了回归保护，避免类似 PR #26521 和 #26655 的 bug 再次逃逸到生产环境。新增的测试方法被注册到 base-b stage，将在 4-gpu-b200 和 1-gpu-large 配置上运行。
- 风险标记：暂无

关联脉络

- PR #26521 fix: copy seq_lens in TRTLLM MHA draft decode cuda graph capture: 本 PR 新增的测试专门覆盖 PR #26521 修复的 CG capture 路径，并通过 bug 注入验证测试有效性。
- PR #26655 Fix TRTLLM MHA draft decode cache seqlens replay: 本 PR 新增的测试覆盖 PR #26655 修复的 CG replay 路径，但测试说明当前单次 replay 尚未暴露该 bug。
- PR #26658 test: strengthen CG-replay coverage with prod-fill padding, metadata invariants, and pad-ratio sweep: 本 PR 依赖 PR #26658 引入的 metadata-invariant 辅助函数。