PR #33607 完整报告

sgl-project/sglang

[ci] add qwen 3.5 mtp + replayssm + flashinfer gdn test

合并时间：2026-08-05 09:08

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33607

执行摘要

- 一句话：新增 Qwen3.5 MTP+ReplaySSM+FlashInfer GDN 端到端测试
- 推荐动作：值得快速浏览（约 5 分钟）。本 PR 是纯测试补强，虽无源码改动，但类 docstring 清晰记录了两种设计决策：一是用 --mamba-ssm-dtype bfloat16 覆盖 ReplaySSM 的 float32 默认状态，二是用三个显式后端参数钉住 FlashInfer，防止自动选择漂移。这种『显式钉住后端 + 覆盖默认 dtype』的测试写法，适合作为新增 spec-verify 协议回归测试的模板。不建议精读源码，因为不涉及运行时逻辑。

功能与动机

PR body 为模板内容，未给出文字动机，动机主要体现在新增测试类的 docstring 中：需要为 MTP 与 ReplaySSM spec-verify 折叠协议的组合提供端到端回归覆盖，并显式钉住 FlashInfer GDN（bf16-state）内核栈——因为 --enable-linear-replayssm-spec 默认会把线性 SSM 状态 dtype 设为 float32，而测试希望用 bf16 状态验证；同时用三个后端标志防止线性注意力自动选择默认值漂移导致测试失去针对性。

实现拆解

1. CI 预估时长调整：在 test/registered/models_e2e/test_qwen35_fp4_mtp.py 中将 register_cuda_ci(est_time=...) 从 400 上调到 800 秒。原因：文件内新增第二个端到端用例后，测试文件在 base-c 阶段 4-gpu-b200 runner 上的总运行时间预计翻倍，需要让 CI 调度器对时长预估更准确。
2. 复用既有 MTP 启动参数：保持 MTP_BASE_ARGS 不变（tp-size 4、NEXTN 投机算法 3 步、topk 1、modelopt_fp4 量化、trtllm_mha 注意力后端、mamba-ssm-dtype bfloat16 等），新测试类直接继承该参数列表，避免重复维护。
3. 新增 TestQwen35FP4MTPReplaySSM 测试类：setUpClass 在 MTP_BASE_ARGS 基础上追加 --enable-linear-replayssm-spec 与 --linear-attn-decode/prefill/verify-backend flashinfer 三个标志。设计意图见类 docstring：ReplaySSM 折叠协议默认可能把线性 SSM 状态 dtype 切到 float32，而 MTP_BASE_ARGS 中的 --mamba-ssm-dtype bfloat16 显式覆盖回 bf16；三个后端参数把 decode/prefill/verify 阶段的线性注意力实现锁定为 FlashInfer，防止自动选择默认值漂移。
4. 复用校验逻辑：test_gsm8k 调用模块级 _run_mtp_gsm8k，同时断言 gsm8k 准确率达到 0.95 阈值、平均投机接受长度 avg_spec_accept_length 大于 3.3，验证 ReplaySSM 协议下投机解码仍保持足够的接受率。

5. 测试配套与配置：不修改任何运行时源码，无 schema、部署或文档配套改动；文件内唯一配置改动是 est_time 注册元数据。

关键文件：

- test/registered/models_e2e/test_qwen35_fp4_mtp.py（模块 端到端测试；类别 test；类型 test-coverage；符号 TestQwen35FP4MTPReplaySSM, setUpClass, tearDownClass, test_gsm8k）：本 PR 唯一变更文件：新增 TestQwen35FP4MTPReplaySSM 端到端测试类，并调整 register_cuda_ci 的 est_time，是 CI 矩阵扩展的核心载体。

关键符号：TestQwen35FP4MTPReplaySSM, setUpClass, tearDownClass, test_gsm8k

评论区精华

该 PR 未产生实质 review 讨论（0 条 review 评论、0 条 review threads）。comments_count 中的 2 条互动均为 CI 操作：作者发起 /rerun-test test/registered/models_e2e/test_qwen35_fp4_mtp.py，github-actions[bot] 随后报告 4-gpu-b200 上该测试通过（✅）。因此本 PR 的决策主要由作者单方面完成，包括 est_time 上调与后端参数显式钉住，未见他人对设计提出异议或补充。

- 暂无高价值评论线程

风险与影响

- 风险：风险集中在 CI 与测试有效性层面： 1) est_time 从 400 翻倍到 800 秒，若新用例实际耗时低于预估会浪费 CI 调度窗口；若高于 800 秒，base-c 阶段可能超时。 2) 新用例显式依赖 FlashInfer 的 linear-attn decode/prefill/verify 三个后端，若 FlashInfer 线性注意力内核在 bf16 状态或 Blackwell 架构上存在隐藏缺陷，该测试会成为后端的强耦合回归点。 3) 用例仅覆盖 gsm8k 单数据集与单模型（Qwen3.5-397B-A17B-NVFP4），对 ReplaySSM 与 FlashInfer 组合的泛化性验证有限。 4) 该测试为端到端重型用例（4 GPU B200），相比仓库最近把量化 e2e 矩阵换成 layer-level 单测的趋势（如 PR #33596、#33611），本 PR 反向增加 e2e 资源消耗，可能影响 base-c CI 总时长预算。
  - 影响：对用户与运行时无影响（纯测试文件变更）。对系统 CI 的影响：base-c 阶段 4-gpu-b200 runner 上该测试文件预估耗时从 400 秒增至 800 秒，B200 资源占用翻倍，与近期 PR #33586 精简 B200 冗余注册、缩减约 27 分钟的方向存在张力。对团队的影响：新增的 TestQwen35FP4MTPReplaySSM 为 Qwen3.5 的 NEXTN+ReplaySSM+FlashInfer GDN 组合提供了端到端回归护栏，后续任何修改 --enable-linear-replayssm-spec 默认 dtype 或线性注意力后端自动选择逻辑的 PR 都会被该测试拦住。
  - 风险标记：CI 时长预估翻倍，依赖 FlashInfer 线性注意力后端，单模型单数据集覆盖

关联脉络

- PR #33586 [CI] Trim redundant B200 test registrations: 同一文件 test/registered/models_e2e/test_qwen35_fp4_mtp.py 的 CI 注册调整，本 PR 新增用例并上调 est_time，需要与 33586 的 B200 时长精简目标协调。

- PR #33611 [Test] Replace NVFP4 MoE runner backend e2e matrix with a layer-level unit test: 同样针对 NVFP4 量化模型，但 33611 是把 *e2e* 矩阵替换为 layer-level 单测以提速 CI，本 PR 反向增加 *e2e* 用例，两者体现 CI 策略上的不同取向。