

PR #27149 完整报告

sgl-project/sglang

[AMD] [CI] Add dsv4 accuracy PR gate to pr-test-amd-rocm720

合并时间: 2026-06-12 14:42

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27149>

执行摘要

- 一句话: 为 AMD CI 添加 DSV4 精度测试门控
- 推荐动作: 值得阅读, 展示了 CI 门控设计模式 (按事件类型过滤、组合 cron 与 PR 触发、条件跳过测试), 适合作为 CI 配置参考。

功能与动机

PR body 说明: 'Adds ROCm 7.2 DeepSeek-V4 accuracy coverage to pr-test-amd-rocm720.yml', 目的是在 AMD ROCm 平台上为 DeepSeek-V4 模型建立持续集成精度验证, 防止回归并确保 AMD 硬件上的推理正确性。

实现拆解

1. 修改 CI workflow 触发器 (.github/workflows/pr-test-amd-rocm720.yml) : 启用 pull_request 并限制路径; 新增 0 */6 * * * cron 用于只运行 dsv4 精度作业; 在 workflow_dispatch 的 stage 选项中添加 dsv4-flash-fp4-fp8-amd-rocm720 和 dsv4-pro-fp4-amd-rocm720。
2. 调整 call-gate 和 check-changes 条件: call-gate 添加 if: github.event.schedule != '0 */6 * * *'; check-changes 中 set-run-all-tests 步骤增加 || "\${{ github.event_name }}" == "schedule"。
3. 添加 dsv4 精度测试 job: 基于现有 dsv4 测试 job, 在 PR 和 6h cron 中通过依赖控制仅运行 dsv4 作业; 设置 SGLANG_DSV4_ACCURACY_ONLY=1 环境变量使性能测试跳过。
4. 更新四个测试脚本的环境变量和 skip 装饰器: 将 SGLANG_HACK_FLASHMLA_BACKEND 改为 unified_kv_triton; 在 test_b_perf_8k_1k 方法上添加 @unittest.skipIf。
5. 配置 continue-on-error: PR 触发时 false, cron 触发时 true。

关键文件:

- .github/workflows/pr-test-amd-rocm720.yml (模块 CI 配置; 类别 infra; 类型 infrastructure) : CI workflow 核心变更, 定义了 DSV4 精度测试的触发条件和 dsv4 作业集成。
- test/registered/amd/test_deepseek_v4_flash_fp4.py (模块 精度测试; 类别 test; 类型 test-coverage) : DSV4 Flash FP4 精度测试脚本, 包含环境变量调整和 skipIf 装饰器。

- test/registered/amd/test_deepseek_v4_flash_fp8.py (模块 精度测试; 类别 test; 类型 test-coverage) : DSV4 Flash FP8 精度测试脚本, 变更同 flash_fp4。
- test/registered/amd/test_deepseek_v4_pro_fp4.py (模块 精度测试; 类别 test; 类型 test-coverage) : DSV4 Pro FP4 精度测试脚本, 变更同前。
- test/registered/amd/test_deepseek_v4_pro_fp8.py (模块 精度测试; 类别 test; 类型 test-coverage) : DSV4 Pro FP8 精度测试脚本, 变更同前。

关键符号: TestDeepseekV4Fp4.test_b_perf_8k_1k,
TestDeepseekV4Fp8.test_b_perf_8k_1k, TestDeepseekV4ProFp4.test_b_perf_8k_1k,
TestDeepseekV4ProFp8.test_b_perf_8k_1k

关键源码片段

test/registered/amd/test_deepseek_v4_flash_fp4.py

DSV4 Flash FP4 精度测试脚本, 包含环境变量调整和 skipIf 装饰器。

```
# test/registered/amd/test_deepseek_v4_flash_fp4.py ( 关键变更 )
# 公共环境变量, AMD ROCm 7.2 专用配置
COMMON_ENV_VARS = {
    # ... 其他环境变量 ...
    # 变更前: 'SGLANG_HACK_FLASHMLA_BACKEND': 'triton'
    # 变更后: 统一使用 unified_kv_triton 后端
    'SGLANG_HACK_FLASHMLA_BACKEND': 'unified_kv_triton',
    # ... 其余环境变量保持不变 ...
}

class TestDeepseekV4Fp4(CustomTestCase):
    # ... setUpClass 与 tearDownClass 不变 ...

    # 精度测试 (直接执行, 不跳过)
    def test_a_gsm8k(self):
        # 运行 GSM8K 评估并断言准确率 > 0.91
        ...

    # 性能测试: 当 SGLANG_DSV4_ACCURACY_ONLY=1 时跳过
    @unittest.skipIf(
        os.environ.get('SGLANG_DSV4_ACCURACY_ONLY') == '1',
        'SGLANG_DSV4_ACCURACY_ONLY=1: accuracy-only run (skipping perf)',
    )
    def test_b_perf_8k_1k(self):
        # 运行基准测试, 输入 8K token 输出 1K token
        ...
```

评论区精华

此 PR 没有实质性的 review 讨论, 仅有一个自动评论和 HaiShaw 的批准。审核者直接批准, 表明变更清晰且风险低。

- 暂无高价值评论线程

风险与影响

- 风险：未发现明显风险。环境变量 `SGLANG_DSV4_ACCURACY_ONLY` 可能在某些 PR 中跳过性能测试，导致性能回归泄漏；FlashMLA 后端切换可能需要同步其他 CI 配置；新 cron 作业可能增加 ROCm CI 资源消耗。
- 影响：对用户无影响；对开发团队：增强了 AMD 平台 DeepSeek-V4 的 CI 覆盖，但也增加了 CI 总时间和资源占用；对 CI 流程引入了更精细的触发条件控制。
- 风险标记：新环境变量跳过性能测试，FlashMLA 后端变更，CI 资源消耗增加，PR 触发条件可能遗漏非标准路径

关联脉络

- PR #27994 Remove outdated patch: 同为 AMD CI 维护，本 PR 在 27994 基础上新增 dsv4 门控，共同推进 AMD CI 稳定性。
- PR #27999 [AMD] Pin maturin<1.14 to fix ROCm image build failure: 同为 AMD CI 维护，本 PR 与 27999 均涉及 AMD ROCm CI 流程优化。