PR #24825 完整报告

sgl-project/sglang

[AMD] DSv4 nightly hotfix + schedule-aware --continue-on-error in AMD CI

合并时间：2026-05-11 12:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24825

执行摘要

- 一句话：修复 AMD CI DSv4 参数回退和 cron 续跑问题
- 推荐动作：推荐 AMD 和 CI 维护者关注本次 continue-on-error 条件的设计模式（github.event_name == 'schedule'），以及跨分支配置兼容性处理。对于其他硬件后端，可借鉴类似的条件感知方式。本 PR 改动直观、测试充分，值得精读。

功能与动机

自 2026-05-08 起 AMD CI nightly 因 #23882 分支重命名而频繁失败。PR #23882 将 attention-backend 值 compressed 改为 dsv4，环境变量 SGLANG_REASONING_EFFORT 改为 SGLANG_DSV4_REASONING_EFFORT，但 DSv4 镜像未同步该变更，导致 argparse 拒绝新参数。同时 cron 触发时因 inputs 为空导致 continue-on-error 未生效，首个失败即终止整体测试。

实现拆解

1. 回退 DSv4 测试参数：修改 test/registered/amd/test_deepseek_v4_flash_fp4.py 等四个文件的 other_args 中 --attention-backend 从 dsv4 恢复为 compressed，确保 DSv4 镜像能够识别；环境变量保留新名称 SGLANG_DSV4_REASONING_EFFORT（最终提交确认镜像已支持）。同时将 FP4/FP8 文件重命名以包含 flash 标记，统一命名规范。
2. 修复continue-on-error条件：在nightly-test-amd-rocm720.yml、nightly-test-amd.yml、pr-test-amd-rocm720.yml 中，将所有 run_suite.py 调用的条件从 ${{ inputs.continue_on_error && '--continue-on-error' || '' }} 替换为 ${{ (github.event_name == 'schedule' || inputs.continue_on_error) && '--continue-on-error' || '' }}，确保 cron 调度时自动启用续跑，同时保留手动触发时的控制权。
3. 重命名测试文件：将 test_deepseek_v4_fp4.py 和 test_deepseek_v4_fp8.py 分别重命名为 test_deepseek_v4_flash_fp4.py 和 test_deepseek_v4_flash_fp8.py，匹配 Pro 和 NV 变体的命名习惯，文件内注册类名不变，无需调整 run_suite.py。

关键文件：

- .github/workflows/nightly-test-amd-rocm720.yml（模块 CI 配置；类别 infra；类型 infrastructure）：最关键的修复：修正 continue-on-error 条件，确保 cron 调度时自动启用续跑，影响所有 nightly 作业

- .github/workflows/nightly-test-amd.yml（模块 CI 配置；类别 infra；类型 infrastructure）：与 nightly-test-amd-rocm720.yml 相同的 continue-on-error 修复，覆盖 AMD nightly 所有作业
- test/registered/amd/test_deepseek_v4_flash_fp4.py（模块 DSv4 测试；类别 test；类型 rename-or-move）：展示 attention-backend 回退和文件重命名，确保镜像兼容
- test/registered/amd/test_deepseek_v4_flash_fp8.py（模块 DSv4 测试；类别 test；类型 rename-or-move）：与 flash_fp4 相同，回退 attention-backend 并重命名文件
- .github/workflows/pr-test-amd-rocm720.yml（模块 CI 配置；类别 infra；类型 infrastructure）：与 nightly 相同的 continue-on-error 修复，应用于 PR 测试
- test/registered/amd/test_deepseek_v4_pro_fp4.py（模块 DSv4 测试；类别 test；类型 test-coverage）：Pro 变体同样回退 attention-backend 参数
- test/registered/amd/test_deepseek_v4_pro_fp8.py（模块 DSv4 测试；类别 test；类型 test-coverage）：Pro FP8 变体回退 attention-backend 参数

关键符号：未识别

关键源码片段

test/registered/amd/test_deepseek_v4_flash_fp4.py

展示 attention-backend 回退和文件重命名，确保镜像兼容

```
# 公共环境变量：使用新名称 SGLANG_DSV4_REASONING_EFFORT（镜像已支持）
COMMON_ENV_VARS = {
    "SGLANG_DSV4_REASONING_EFFORT": "max", # 推理努力程度
    # ... 其他变量
}

class TestDeepseekV4Fp4(CustomTestCase):
    @classmethod
    def setUpClass(cls):
        cls.model = DEEPSEEK_V4_FP4_MODEL_PATH
        cls.base_url = DEFAULT_URL_FOR_TEST

        env = os.environ.copy()
        env.update(COMMON_ENV_VARS)
        env.update(FP4_ENV_VARS)

        other_args = [
            "--trust-remote-code",
            "--tp", "8",
            "--disable-radix-cache",
            "--attention-backend",
            "compressed", # 使用旧名称以确保 DSv4 镜像能够识别
            "--max-running-requests", "256",
            "--page-size", "256",
            "--chunked-prefill-size", "8192",
            "--disable-shared-experts-fusion",
```

```
            "--tool-call-parser", "deepseekv4",
            "--reasoning-parser", "deepseek-v4",
        ]

        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=SERVER_LAUNCH_TIMEOUT,
            other_args=other_args,
            env=env,
        )
```

评论区精华

无显著讨论：仅有一个来自 HaiShaw 的批准，无 review 评论。PR body 详细说明了问题分析和解决方案的权衡。最后一个提交将环境变量恢复为新名称，表明镜像已经兼容，但 attention-backend 仍使用旧名称，确保最大兼容性。

- 暂无高价值评论线程

风险与影响

- 风险：兼容性风险：恢复使用 compressed 依赖主分支的别名支持，若未来移除别名将失效；但临时方案且有明确回退计划。continue-on-error 条件修改可能影响手动触发时的行为，但保留了 inputs.continue_on_error 的控制，风险可控。测试文件重命名不影响注册，因为 suite 注册在文件内。总体风险较低。
- 影响：对 AMD CI 运维者：修复后 nightly 和 PR 测试可稳定运行，减少因参数不匹配和 fail-fast 导致的误报。对其他平台无影响。对 SGLang 项目：提升了 CI 对跨分支配置不同步的容错能力，并提供了可复用的 schedule 感知条件模式。
- 风险标记：兼容性依赖别名，cron vs manual 条件判断

关联脉络

- PR #23882 Deepseek V4 rename (compressed -> dsv4): 是导致 DSv4 测试参数不兼容的直接原因，本 PR 的部分变更正是对其重命名的回退兼容处理