

# PR #24203 完整报告

sgl-project/sglang

[AMD] Deepseek v4 Flash / Pro nightly tests for MI35x ROCm 7.2

合并时间: 2026-05-04 15:02

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/24203>

## 执行摘要

- 一句话: 为 MI35x ROCm 7.2 添加 DSv4 Flash/Pro 的 nightly 测试
- 推荐动作: 建议阅读工作流中关于避免重建 sglang/aiter 的技巧 (`--skip-sglang-build` 和容器路径隔离), 以及动态解析最新 DSv4 镜像标签的做法。测试文件本身的模式可以作为后续添加类似模型的模板。

## 功能与动机

确保 DeepSeek-V4 系列模型在 AMD ROCm 7.2 平台的长期正确性和性能, 防止回归。

## 实现拆解

1. 新增四个测试文件: `test_deepseek_v4_fp4.py`、`test_deepseek_v4_fp8.py` (Flash 模型), `test_deepseek_v4_pro_fp4.py`、`test_deepseek_v4_pro_fp8.py` (Pro 模型)。每个测试类共享一个服务器进程, 依次运行 GSM8K 精度评估和 `bench_one_batch_server` 性能基准。
2. 修改 `nightly-test-amd-rocm720.yml` 工作流, 添加两个新 job (flash 和 pro), 每个 job 包含: 动态解析 DSv4 容器标签 (从 Docker Hub 获取最新 `rocm720-mi35x-*-DSv4` 镜像), 启动容器时传递 `--skip-sglang-build` 和 `--skip-aiter-build` 以保留容器内预置的运行时。
3. 环境变量配置集中管理, 区分公共通用变量和每个变体特有的变量 (如 FP4/FP8 开关)。
4. 注册 nightly 测试套件并设置超时 (Flash 3600s, Pro 5400s), 精度阈值分别为 0.91 (Flash) 和 0.92 (Pro)。

关键文件:

- `test/registered/amd/test_deepseek_v4_pro_fp4.py` (模块 模型测试; 类别 `test`; 类型 `test-coverage`; 符号 `TestDeepseekV4ProFp4`, `setUpClass`, `tearDownClass`, `test_a_gsm8k`): 测试用例模板, 覆盖 Pro FP4 变体的精度和性能测试, 包含环境变量、启动参数、超时等关键配置。
- `test/registered/amd/test_deepseek_v4_pro_fp8.py` (模块 模型测试; 类别 `test`; 类型 `test-coverage`; 符号 `TestDeepseekV4ProFp8`, `setUpClass`, `tearDownClass`, `test_a_gsm8k`): Pro FP8 变体的对称测试, 规则与 FP4 类似但环境变量不同。
- `test/registered/amd/test_deepseek_v4_fp4.py` (模块 模型测试; 类别 `test`; 类型 `test-coverage`; 符号 `TestDeepseekV4Fp4`, `setUpClass`, `tearDownClass`, `test_a_gsm8k`): Flash FP4 变体的测试, 与 Pro 相比超时更短、精度阈值略低。

- test/registered/amd/test\_deepseek\_v4\_fp8.py (模块 模型测试; 类别 test; 类型 test-coverage; 符号 TestDeepseekV4Fp8, setUpClass, tearDownClass, test\_a\_gsm8k) : Flash FP8 变体的测试。
- .github/workflows/nightly-test-amd-rocm720.yml (模块 CI 工作流; 类别 infra; 类型 infrastructure) : CI 工作流更新, 新增两个 nightly job 并引入 DSv4 镜像动态标签解析和运行时保留技巧。

关键符号: setUpClass, tearDownClass, test\_a\_gsm8k, test\_b\_perf\_8k\_1k

## 评论区精华

无实质性 review 讨论, 仅有一位 maintainer 直接批准。HaiShaw 在合并后留言希望改进 nightly 测试的信号, 但未在 PR 内展开讨论。

- 改进 nightly 测试信号 (other): 未在 PR 内讨论, 可能需要后续 PR 处理。

## 风险与影响

- 风险: 低风险。变更仅涉及测试文件和 CI 工作流, 不影响任何生产代码。主要风险在于: 若测试机环境配置不正确 (如缺少 ROCm 700A 或 tilelang 依赖), 可能导致假失败; DSv4 容器标签解析逻辑依赖 Docker Hub 返回次序, 若镜像策略改变可能失败。但这些属于基础设施稳定性问题。
- 影响: 对用户无直接影响。对团队影响: 增加了对 AMD 平台 DeepSeek-V4 模型的回归保护, 有利于及早发现精度或性能退化。CI 执行时间延长 (两个 job 各约 1-2 小时)。
- 风险标记: 依赖外部 Docker Hub 标签解析, 测试环境依赖 ROCm 700A 和 tilelang

## 关联脉络

- PR #23538 [NPU] Fix Z-Image negative-branch rotary embeddings for CFG: 同为 AMD 平台 bugfix, 体现对 AMD 测试的持续加强。
- PR #24230 [pd]: (Bug Fix) Incorrect out\_cache\_loc slicing in prepare\_for\_prebuilt: 同为 AMD/ROCm 相关的 bugfix, 说明平台测试的重要性。
- PR #24270 [codex] Add official diffusion GT workflow mode: 同为 CI 基础设施增强, 增加强制 GT 模式。