PR #25260 完整报告

sgl-project/sglang

[AMD][CI] Register Eagle constrained decoding test

合并时间：2026-05-17 14:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25260

执行摘要

- 一句话：AMD CI 注册 EAGLE 约束解码测试
- 推荐动作：该 PR 变更简单，但体现了跨平台 CI 覆盖的重要实践：为 AMD 专用 runner 添加上游已有 CUDA 测试时，应根据实际运行时长合理调整 est_time 以优化 CI 分区。可快速合入。

功能与动机

为 AMD/ROCm 平台增加 EAGLE 约束解码测试的 CI 覆盖，以便在 AMD 环境上验证 EAGLE、EAGLE V2 以及 xgrammar 支持的约束生成路径，从而提前发现 AMD 相关的回归问题。

实现拆解

1. 导入新增：在 test/registered/spec/eagle/test_eagle_constrained_decoding.py 的导入语句中，将 from sglang.test.ci.ci_register import register_cuda_ci 改为 from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci，使得 register_amd_ci 可用。
2. 注册 AMD CI：在 register_cuda_ci(est_time=116, stage="base-b", runner_config="1-gpu-large") 之后新增一行 register_amd_ci(est_time=165, stage="stage-b", runner_config="1-gpu-large-amd")，将该测试注册到 AMD CI 套件的 stage-b 阶段，使用 1-gpu-large-amd runner 配置，估计耗时 165 秒（基于 PR body 中 AMD 实际运行约 165 秒的数据）。
3. 无其他改动：测试类、测试方法、测试逻辑均保持不变，仅 CI 注册区域变化。

关键文件：

- test/registered/spec/eagle/test_eagle_constrained_decoding.py（模块 推测解码；类别 test；类型 test-coverage）：唯一变更文件，添加了 AMD CI 注册函数调用和导入更新。

关键符号：未识别

关键源码片段

test/registered/spec/eagle/test_eagle_constrained_decoding.py

唯一变更文件，添加了 AMD CI 注册函数调用和导入更新。

```
import unittest
```

```
from sglang.srt.environ import envs
from sglang.srt.utils import kill_process_tree
# 同时导入 CUDA 和 AMD 的 CI 注册函数
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci
from sglang.test.kits.json_constrained_kit import JSONConstrainedMixin
from sglang.test.kits.regex_constrained_kit import RegexConstrainedMixin
from sglang.test.test_utils import (
    DEFAULT_DRAFT_MODEL_EAGLE,
    DEFAULT_TARGET_MODEL_EAGLE,
    DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    DEFAULT_URL_FOR_TEST,
    CustomTestCase,
    popen_launch_server,
)

# 先注册 CUDA CI（原有）
register_cuda_ci(est_time=116, stage="base-b", runner_config="1-gpu-large")
# 新增 AMD CI 注册，est_time 基于实际 AMD 运行耗时 165s 设置
register_amd_ci(est_time=165, stage="stage-b", runner_config="1-gpu-large-amd")


class TestEagleConstrainedDecoding(
    CustomTestCase, RegexConstrainedMixin, JSONConstrainedMixin
):
    # ... 测试类定义保持不变 ...
```

评论区精华

- AI 助手评论：gemini-code-assist[bot] 指出初始 est_time=116 对 AMD 运行时长估计不足，建议改为 165 秒以优化 CI 分区效率，该建议已被采纳。
- HaiShaw 审批通过：无额外讨论。
  - est_time 调整以匹配 AMD 实际运行时长 (testing): 已采纳建议，est_time 从 116 改为 165。

风险与影响

- 风险：无技术风险。本 PR 仅添加 CI 注册函数调用，不修改任何运行时代码或测试逻辑。可能的风险是若 AMD 环境上其他测试时限冲突或资源争用导致超时，但 est_time 已基于实际运行结果设置，且 CI 框架可控。
- 影响：
  - 对用户：无直接影响。
  - 对系统：AMD GPU CI 流水线将新增 24 个测试用例（约 165 秒），包含 JSON 约束生成、无效 JSON 处理、OpenAI 兼容 JSON 输出以及正则约束生成等场景。
  - 对团队：AMD 平台回归检测覆盖增强，有助于维护 EAGLE 约束解码在 ROCm 上的兼容性。
  - 风险标记：仅 CI 配置变更，无运行时风险

关联脉络

- PR #25457 [diffusion] add memory-aware component load order: 同为在 AMD CI 中注册测试的 PR，展示了跨平台 CI 覆盖的实践模式。