PR #29234 完整报告

sgl-project/sglang

[AMD] Fix stage-b-test-1-gpu-small-amd-nondeterministic timeout after VLM model swap

合并时间：2026-06-25 18:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29234

执行摘要

- 一句话：提高 AMD VLM CI 测试超时时间
- 推荐动作：快速合并，解决 CI 阻塞。建议跟踪测试耗时，未来可考虑优化 VLM 推理或拆分测试集。

功能与动机

PR #29095 临时禁用了 openbmb/MiniCPM-V-2_6 并替换为 Qwen2.5-VL-3B-Instruct。Qwen2.5-VL 使用动态分辨率，高分辨率 MMMU 图像产生更多视觉 token 和更重的预填充，在 AMD/ROCm 上运行时间超过旧限制 1800 秒。

实现拆解

1. 修改 .github/workflows/pr-test-amd.yml 中对应任务的 timeout-per-file 参数从 1800 改为 3600，timeout-minutes 从 45 改为 75。
2. 在 .github/workflows/pr-test-amd-rocm720.yml 中做相同的超时调整。
3. 无源代码或运行时逻辑变更。

关键文件：

- .github/workflows/pr-test-amd.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：主 CI 工作流，调整超时参数以适配新 VLM 模型。
- .github/workflows/pr-test-amd-rocm720.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：ROCm 7.2 CI 工作流，相同超时调整。

关键符号：未识别

关键源码片段

.github/workflows/pr-test-amd.yml

主 CI 工作流，调整超时参数以适配新 VLM 模型。

```
# .github/workflows/pr-test-amd.yml（片段）
- name: Run test
  timeout-minutes: 75 # 从 45 上调，给新模型充分时间
  run: |
    bash scripts/ci/amd/amd_ci_exec.sh -w "/sglang-checkout/test"
```

```
python3 run_suite.py --hw amd \
  --suite stage-b-test-1-gpu-small-amd-nondeterministic \
  --timeout-per-file 3600  # 从 1800 翻倍，适应 Qwen2.5-VL 的预填充开销
  ${{ needs.check-changes.outputs.continue_on_error == 'true' && '--continue-on-error' || '' }}
  }
```

评论区精华

无 review 评论。PR 由 HaiShaw 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅调整 CI 超时参数，不涉及推理代码或性能。但过长的超时可能掩盖真正的挂起问题，需确保测试确实能完成。
- 影响：仅影响 AMD CI 的 stage-b-test-1-gpu-small-amd-nondeterministic 任务，允许 VLM 评估有更长时间执行。对其他平台或任务无影响。
- 风险标记：仅配置变更

关联脉络

- PR #29095 [AMD] swa0 VLM model from MiniCPM-V-2_6 to Qwen2.5-VL-3B-Instruct: 导致超时问题的根本原因 PR，替换了 VLM 模型。