

PR #30309 完整报告

sgl-project/sglang

[AMD] ci: run multimodal_gen unit suite on AMD

合并时间: 2026-07-08 07:02

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30309>

执行摘要

- 一句话: AMD CI 新增 multimodal_gen 单元测试覆盖
- 推荐动作: 建议合并。这是一个标准的 CI 覆盖增强 PR, 设计清晰、测试已验证通过。后续可以关注排除的测试文件是否能随 ROCm 环境改进而重新启用。

功能与动机

PR 描述指出, multimodal_gen 单元测试套件是便携的 CPU 风格单元测试, 不需要 NVIDIA 硬件, 但之前仅在 CUDA CI 中运行, 导致 AMD CI 存在覆盖缺口 (coverage gap)。本 PR 旨在填补这一缺口, 使 AMD ROCm CI 也运行这些测试。

实现拆解

1. 在 `pr-test-amd.yml` 中添加作业: 新增 multimodal-gen-unit-test-amd 作业, 在 ROCm 7.0.0 容器中运行 `run_suite.py --suite unit`, 并通过 `-k "not ltx2_vae_channels_last"` 排除一个 CUDA 特定测试文件。该作业被加入 `target_stage_select` 下拉选项和终点门控。
2. 在 `pr-test-amd-rocm720.yml` 中添加作业: 类似地, 添加 multimodal-gen-unit-test-amd-rocm720 作业, 在 ROCm 7.2.0 环境中运行相同套件, 也排除同一测试文件。
3. 更新 CI 覆盖报告: 在 `scripts/ci/utlis/ci_coverage_report.py` 中, 将 `unit` 子目录的后端映射从 ("CUDA",) 改为 ("CUDA", "AMD"), 并添加若干重构后子目录的映射规则 (如 `single_test_file`、`unit/realtime` 等), 确保覆盖报告准确统计。

关键文件:

- `.github/workflows/pr-test-amd.yml` (模块 AMD 工作流; 类别 infra; 类型 infrastructure) : 在 ROCm 7.0.0 CI 中添加了 multimodal-gen-unit-test-amd 作业, 是本次变更的核心
- `.github/workflows/pr-test-amd-rocm720.yml` (模块 AMD 工作流; 类别 infra; 类型 infrastructure) : 在 ROCm 7.2.0 CI 中镜像添加了对应作业, 确保两个 ROCm 版本都覆盖
- `scripts/ci/utlis/ci_coverage_report.py` (模块 覆盖报告; 类别 infra; 类型 infrastructure) : 更新了 CI 覆盖报告, 将 `unit` 子目录映射为 CUDA 和 AMD, 并添加了重构后子目录的映射规则

关键符号: 未识别

关键源码片段

scripts/ci/utils/ci_coverage_report.py

更新了 CI 覆盖报告，将 unit 子目录映射为 CUDA 和 AMD，并添加了重构后子目录的映射规则

```
# 后端映射字典，指示每个子目录在哪些后端上运行
_MM_GEN_SUBDIR_BACKENDS = {
    "server": ("CUDA",),
    "server/amd": ("AMD",),
    "server/musa": ("MUSA",),
    "layers": ("CUDA",),
    "unit": ("CUDA", "AMD"), # 现在 AMD CI 也运行 unit 套件
    "cli": ("CUDA",),
    "manual": ("CUDA",),
    # 以下为重构后的子目录映射，此前无规则导致被完全丢弃
    "single_test_file": ("CUDA",),
    "single_test_file/component_accuracy": ("CUDA",),
    "unit/realtime": ("CUDA",),
    "unit/sana_wm": ("CUDA",),
    "unit/progressive_resolution": ("CUDA",),
    "unit/musa/layers": ("MUSA",),
}
```

评论区精华

本 PR 无实质性 review 讨论。合并者 HaiShaw 直接批准，未留下评论。

- 暂无高价值评论线程

风险与影响

- 风险：本 PR 仅涉及 CI 配置变更，不修改任何源代码，回归风险极低。但存在以下注意事项：
 - 排除了 ltx2_vae_channels_last.py，该文件中的测试在 ROCm 上因内存格式断言而失败，需关注后续 ROCm 补齐该功能。
 - 新增作业会增加 CI 运行总时长约 2 分钟，但尚在可接受范围。
 - 覆盖报告更新依赖正确的子目录映射，若未来目录结构再调整，需同步更新映射。
 - 影响：对用户无直接影响。对开发团队，AMD CI 的测试覆盖得到提升，可更早发现便携性问题。CI 总运行时间略有增加，但影响极小。此外，覆盖报告将更准确地反映 AMD 的测试覆盖范围。
- 风险标记：排除 CUDA 特定测试，CI 时间增加，覆盖报告依赖目录结构

关联脉络

- PR #30207 [AMD] Register 2 hardware-agnostic 1-GPU PR tests for AMD CI: 同为将便携测试纳入 AMD CI 的系列 PR

- PR #30290 [AMD] Register 5 CI-verified 1-GPU kernel/attention unit tests for AMD
PR CI: 继续扩展 AMD CI 覆盖范围