PR #23944 完整报告

sgl-project/sglang

[AMD] Fix CI test_diffusion_generation[flux_2_image_t2i_2_gpus]

合并时间：2026-04-28 23:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23944

执行摘要

本 PR 将两个 AMD CI 工作流中扩散模型测试步骤的超时时间统一提升至 150 分钟，解决了因 HuggingFace 模型下载缓慢导致的频繁超时问题。变更仅涉及 CI 配置，无代码或测试逻辑改动。

功能与动机

AMD CI 的 multimodal-gen-test-2-gpu-amd 任务执行扩散模型测试（如 FLUX.1-dev）时，需要从 HuggingFace Hub 下载约 39 个文件。在某些网络条件下，每个文件下载需约 4 分钟，90 分钟的步骤超时在下载完成前就被触发，导致任务失败（例如 此运行）。相比之下，NVIDIA 等效工作流使用 240 分钟超时，AMD 工作流的时间预算严重不足。

实现拆解

分别在两个工作流文件中修改同一步骤的 timeout-minutes 值：

1. .github/workflows/pr-test-amd.yml（第 807 行附近）：timeout-minutes: 90 → timeout-minutes: 150
2. .github/workflows/pr-test-amd-rocm720.yml（第 751 行附近）：timeout-minutes: 80 → timeout-minutes: 150

两个文件统一为 150 分钟，为慢速下载提供约 1.7 倍于原值的时间，同时仍低于 NVIDIA 的 240 分钟。

评论区精华

无 review 评论。PR 由 bingxche 直接批准，无争议。

风险与影响

- 风险：极低。仅修改 CI 超时配置，不涉及代码或测试逻辑。
- 影响：消除 AMD CI 中因下载超时导致的假阴性失败。对其他测试步骤、NVIDIA CI 或用户无影响。
- 建议改进：可考虑增加模型缓存或预下载步骤，从根本上减少对网络下载的依赖。

关联脉络

本 PR 与近期其他扩散模型相关的 PR（如 PR#23836 修改默认种子）同属扩散模型 CI 稳定性改进工作，但无直接代码相互依赖。