PR #23045 完整报告

sgl-project/sglang

[AMD] Fix AMD Multimodal Test - skip nvfp4 tests

合并时间：2026-04-18 00:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23045

执行摘要

- 一句话：在 AMD ROCm 平台上跳过 ModelOpt FP8 和 NVFP4 量化测试，修复 CI 失败。
- 推荐动作：该 PR 变更简单直接，适合快速了解如何通过平台检测调整测试覆盖。值得关注的设计决策是选择在测试配置层而非运行时处理硬件不兼容问题，这降低了核心代码复杂度，但可能牺牲测试完整性。建议结合 review 中的讨论，思考未来如何更精细地管理跨平台测试策略。

功能与动机

根据 PR body 描述，在 AMD ROCm 平台上运行 ModelOpt 量化测试会失败：FP8 测试因 torch._scaled_mm 返回 HIPBLAS_STATUS_NOT_SUPPORTED 而失败；NVFP4 测试需要 CUDA 专有的 flashinfer 或 sgl_kernel FP4 内核。因此，需要跳过这些测试以避免 CI 失败。

实现拆解

1. 测试配置调整：修改 python/sglang/multimodal_gen/test/server/gpu_cases.py 文件，在定义 ONE_GPU_CASES_C 测试用例列表时，新增平台条件判断。
2. 条件逻辑实现：使用 current_platform.is_hip() 检测当前是否为 AMD HIP 平台。若是，则将 ONE_GPU_CASES_C 设为空列表，跳过所有 ModelOpt 测试；否则，保留原有的 6 个测试用例（包括 FP8 和 NVFP4）。
3. 注释补充：在代码中添加注释，说明跳过原因：FP8 需要 torch._scaled_mm（ROCm 不支持），NVFP4 需要 CUDA 专有内核。
4. 无其他配套改动：本次变更仅涉及测试配置文件，没有修改源码、部署脚本或文档。

关键文件：

- python/sglang/multimodal_gen/test/server/gpu_cases.py（模块 多模态测试；类别 test；类型 test-coverage）：这是唯一被修改的文件，包含了扩散模型多模态测试的 GPU 用例配置，通过条件逻辑控制 ModelOpt 量化测试的启用，直接影响 CI 执行范围。

关键符号：未识别

关键源码片段

python/sglang/multimodal_gen/test/server/gpu_cases.py

这是唯一被修改的文件，包含了扩散模型多模态测试的 GPU 用例配置，通过条件逻辑控制 ModelOpt 量化测试的启用，直接影响 CI 执行范围。

```
# Skip all ModelOpt tests on AMD: FP8 requires torch._scaled_mm (HIPBLAS_STATUS_NOT_
SUPPORTED
# on ROCm), NVFP4 requires flashinfer or sgl_kernel FP4 kernels (CUDA-only)
if current_platform.is_hip():
    ONE_GPU_CASES_C = [] # 在 AMD 平台上，将测试用例列表设为空，跳过所有 ModelOpt 测试
else:
    ONE_GPU_CASES_C = [
        _make_modelopt_ci_case(
            "flux1_modelopt_fp8_t2i",
            model_path=DEFAULT_FLUX_1_DEV_MODEL_NAME_FOR_TEST,
            modality="image",
            sampling_params=MODELOPT_T2I_CI_sampling_params,
            extras=["--transformer-path", MODELOPT_FLUX1_FP8_TRANSFORMER],
        ),
        # ... 其他 5 个 ModelOpt 测试用例（包括 FP8 和 NVFP4）保持不变
    ]
```

评论区精华

Review 中，gemini-code-assist[bot] 指出当前实现仅通过 not current_platform.is_hip() 排除 AMD 平台，但 NVFP4 测试在其他非 NVIDIA 平台（如 Ascend NPU 或 CPU）上也可能失败。建议使用更具体的检查（如 NVIDIA CUDA 或 Blackwell 架构支持）来防护。此讨论点出了当前方案可能存在的平台兼容性风险，但 PR 最终被批准合并，未采纳该建议。

- 平台检测条件是否足够精确 (correctness): PR 被批准合并，未采纳该建议，维持原实现。

风险与影响

- 风险：1. 测试覆盖风险：在 AMD 平台上完全跳过 ModelOpt 量化测试，可能导致相关功能在 AMD 环境下的回归问题无法被及时发现。 2. 平台兼容性风险：如 review 评论所述，当前条件仅排除 AMD，但 NVFP4 测试在其他非 NVIDIA 平台（如 Ascend NPU、CPU）上运行时仍会失败，可能导致 CI 中断。 3. 维护风险：硬编码的平台检测逻辑可能随硬件支持变化而过时，需要持续更新。
- 影响：1. 对用户：无直接影响，因为这是内部 CI 测试调整，不改变用户可见功能。 2. 对系统：修复了 AMD CI 失败，提高了 CI 稳定性，但减少了 AMD 平台上的测试覆盖范围。 3. 对团队：简化了 AMD 多模态测试的维护，避免了因不支持特性导致的 CI 噪音，但可能增加跨平台测试一致性的管理复杂度。
- 风险标记：测试覆盖缩减，平台兼容性风险

关联脉络

- PR #23031 Revert "feat: Support MXFP4 quantized dense models on AMD CDNA2/CDNA3 GPUs (#19143)": 同样涉及 AMD 平台上的量化模型支持问题，但该 PR 是回退功能，而本 PR 是调整测试覆盖，两者都反映了 AMD 在量化特性上的兼容性挑战。

- PR #22952 [AMD] Add SGLANG_MORI_MOE_MAX_INPUT_TOKENS to truncate dispatch before MoE.: 同为 AMD 相关 PR，涉及性能优化和环境变量配置，展示了团队对 AMD 平台的持续适配努力。
- PR #23076 [diffusion] CI: fix auto-partition: 同为扩散模型 CI 修复，涉及测试配置调整，与本 PR 在测试基础设施层面有协同。