

PR #23955 完整报告

sgl-project/sglang

[AMD] Add AMD FP8 MLA attention test for Wan2.2-T2V-A14B

合并时间: 2026-05-08 16:03

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/23955>

执行摘要

- 一句话: 为 AMD 添加 Wan2.2 FP8 MLA 注意力测试
- 推荐动作: 值得阅读此 PR 的测试模式, 尤其是 DiffusionServerBase 的使用和 CI 注册方法。但应关注 review 中未采纳的改进点, 在后续测试中进行修正, 以确保参数传递正确。

功能与动机

确保 FP8 MLA 代码路径在 AMD nightly CI 中得到覆盖, 捕获回归问题。该路径由 PR #20319 引入, 缺乏自动化测试。

实现拆解

1. 导入与注册

在 `test/registered/amd/test_wan22_fp8_mla.py` 中导入测试基础设施 (`DiffusionServerBase`、`DiffusionTestCase` 等), 并通过 `register_amd_ci` 将测试注册为 nightly 套件 `nightly-amd-fp8-mla-diffusion`, 预估耗时 3600 秒。

2. 定义测试用例配置

使用 `DiffusionTestCase` 数据结构定义 4 组参数: 模型统一为 `Wan-AI/Wan2.2-T2V-A14B-Diffusers`, 分别设置 1/8 GPU、`--enable-torch-compile` 开关 (注意: 当前版本将 flag 和 value 合并在同一字符串中, 可能引发服务器参数解析错误), 8GPU 时指定 `ulysses_degree=4` 和 `cfg_parallel=True`。所有测试均关闭性能、一致性、模型 API 等检查。

3. 实现测试类与夹具

`TestWan22FP8MLA` 继承 `DiffusionServerBase`, 通过 `pytest.fixture` 注入参数化用例, 并重写 `teardown_class` 以安全清理。

4. 实现测试方法

`test_diffusion_generation` 方法根据用例参数生成生成函数, 调用 `run_and_collect` 执行推理, 最后断言输出非空。

关键文件:

- test/registered/amd/test_wan22_fp8_mla.py (模块 测试; 类别 test; 类型 test-coverage ; 符号 TestWan22FP8MLA, teardown_class, case, test_diffusion_generation) : 唯一 变更文件, 新增整个测试套件, 覆盖 FP8 MLA 注意力路径的 AMD CI 测试。

关键符号: TestWan22FP8MLA.teardown_class, TestWan22FP8MLA.case, TestWan22FP8MLA.test_diffusion_generation

评论区精华

Gemini-code-assist[bot] 提出了 3 类建议:

- 参数拆分: extras 列表中的 "--enable-torch-compile false" 应拆为 ["--enable-torch-compile", "false"], 避免服务器启动时被当作单参数。
- 显式设置 tp_size: 8GPU 用例中, ulysses_degree=4 时建议显式添加 tp_size=2 确保正确的模型划分。

所有建议均未在合并前得到修改或回复, 也未标记为已解决。

- 测试参数和配置的正确性 (correctness): 建议未被采纳 (或未被回复), PR 已合并但参数未修改。

风险与影响

- 风险: 测试有效性风险: 所有用例关闭了性能、一致性等检查, 仅验证输出非空, 可能遗漏精度或性能退化。 环境依赖风险: 测试依赖 `sgl_kernel`, 若未安装则跳过 (通过 `pytest.importorskip`), 可能导致 CI 透传但实际未执行。 参数解析风险: `extras` 中未拆分 flag/value 可能造成服务器启动失败, 但测试仍可能因错误处理而误判为通过。
- 影响: 影响范围: 仅影响 AMD nightly CI, 增加约 1 小时执行时间。 用户影响: 无, 纯测试新增。 团队影响: 为后续 AMD 相关测试提供了可参考的模板 (基于 `DiffusionServerBase` 的注册方式)。
- 风险标记: 缺少正确性验证, 环境依赖 (`sgl_kernel`), 命令行参数解析风险

关联脉络

- 暂无明显关联 PR