PR #26775 完整报告

sgl-project/sglang

fix test cases failed on 5/30 in nightly pipeline

合并时间：2026-06-04 20:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26775

执行摘要

- 一句话：修复 NPU 夜间测试失败问题
- 推荐动作：该 PR 为常规维护性修复，无需精读。但可关注 test_ascend_utils.py 中权重路径和环境变量的管理方式，作为测试基础设施维护的参考。

功能与动机

在 5 月 29 日夜间构建中，之前跳过的测试用例重新执行，但在 5 月 30 日触发的夜间流水线中，有 5 个测试用例由于自身实现问题执行失败。

实现拆解

1. 修正 import 路径：在 test_npu_no_overlap_scheduler.py 中将 run_mmlu_test 的导入路径从 sglang.test.ascend.test_ascend_utils 改为 sglang.test.test_utils，并调整为从同一模块导入多个符号。
2. 更新模型权重路径：在 test_ascend_utils.py 中将 KIMI_VL_A3B_INSTRUCT_WEIGHTS_PATH 的模型权重目录从 Kimi/Kimi-VL-A3B-Instruct 更新为 moonshotai/Kimi-VL-A3B-Instruct。
3. 添加环境变量：在 test_npu_deepep_auto_qwen3_next.py 和 test_npu_deepep_low_latency_qwen3_next.py 的环境变量字典中添加 GDN_ATTN_BACKEND_TRITON=1，以适配产品代码更新。
4. 补充测试框架参数：在 test_ascend_utils.py 的 get_benchmark_args 函数调用中添加 ready_check_timeout_sec=0 参数，解决测试框架未适应代码更新导致失败的问题。

关键文件：

- test/registered/ascend/basic_function/parameter/test_npu_no_overlap_scheduler.py（模块 调度测试；类别 test；类型 test-coverage）：修正了错误的 import 路径，将 run_mmlu_test 从 sglang.test.ascend.test_ascend_utils 改为 sglang.test.test_utils
- python/sglang/test/ascend/test_ascend_utils.py（模块 测试工具；类别 test；类型 test-coverage）：更新了 Kimi-VL 模型权重路径并添加了 ready_check_timeout_sec 参数
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_auto_qwen3_next.py（模块 DeepEP 测试；类别 test；类型 test-coverage）：添加环境变量 GDN_ATTN_BACKEND_TRITON=1 以适配产品代码更新

- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_low_latency_qwen3_next.py（模块 DeepEP 测试；类别 test；类型 test-coverage）：添加环境变量 GDN_ATTN_BACKEND_TRITON=1 以适配产品代码更新

关键符号：未识别

关键源码片段

test/registered/ascend/basic_function/parameter/test_npu_no_overlap_scheduler.py

修正了错误的 import 路径，将 run_mmlu_test 从 sglang.test.ascend.test_ascend_utils 改为 sglang.test.test_utils

```
# test_npu_no_overlap_scheduler.py（变更后）
import unittest

from sglang.test.ci.ci_register import register_npu_ci
from sglang.test.test_utils import CustomTestCase, run_mmlu_test # 统一从 test_utils 导入

register_npu_ci(
    est_time=400,
    suite="nightly-1-npu-a3",
    nightly=True,
)

class TestOverlapSchedule(CustomTestCase):
    # ... 测试方法保持不变
```

python/sglang/test/ascend/test_ascend_utils.py

更新了 Kimi-VL 模型权重路径并添加了 ready_check_timeout_sec 参数

```
# test_ascend_utils.py（变更后片段）
# 模型权重路径常量
KIMI_VL_A3B_INSTRUCT_WEIGHTS_PATH = os.path.join(
    MODEL_WEIGHTS_DIR, "moonshotai/Kimi-VL-A3B-Instruct" # 从 Kimi 改为 moonshotai
)

# 在 get_benchmark_args 函数中调用 benchmark_serving 时补充参数
def get_benchmark_args(...):
    # ... 其他参数
    return benchmark_serving(
        # ...
        max_concurrency=max_concurrency,
        ready_check_timeout_sec=0, # 新增参数，解决测试框架未适配问题
    )
```

test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_auto_qwen3_next.py

添加环境变量 GDN_ATTN_BACKEND_TRITON=1 以适配产品代码更新

```
# test_npu_deepep_auto_qwen3_next.py（变更后环境变量部分）
env={
    "SGLANG_DEEPEP_BF16_DISPATCH": "1",
    "PYTORCH_NPU_ALLOC_CONF": "expandable_segments:True",
    "STREAMS_PER_DEVICE": "32",
    "HCCL_OP_EXPANSION_MODE": "AIV",
    "HCCL_ALGO": "level0:NA;level1:ring",
    "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK": "20",
    "HCCL_BUFFSIZE": "2000",
    "GDN_ATTN_BACKEND_TRITON": "1", # 新增环境变量
    **os.environ,
},
```

评论区精华

只有 gemini-code-assist 机器人自动评论总结变更内容，无人工审查讨论。

- 暂无高价值评论线程

风险与影响

- 风险：所有修改仅涉及测试文件（.py 测试用例和测试工具类），不影响生产代码或模型推理逻辑。风险极低，但修改后的测试仅在 NPU 环境中有效，需确保权重路径和环境变量在对应环境中准确可用。
- 影响：直接影响 Ascend NPU 夜间测试流水线，修复 5 个失败的测试用例。对用户和系统无影响。
- 风险标记：低风险，仅测试变更

关联脉络

- PR #26119 [diffusion] Disagg server args, launch helpers, and warmup utils: 同样修改了测试工具类 sglang/test/ascend/test_ascend_utils.py 中的函数，可能涉及类似的参数调整。