PR #26353 完整报告

sgl-project/sglang

NPU Nightly Pipeline Skip Test Case Adaptation and Recovery Testing

合并时间：2026-05-29 09:29

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26353

执行摘要

- 一句话：恢复并适配 NPU Nightly 跳过的测试用例
- 推荐动作：建议合并。该 PR 恢复了重要的 nightly 测试覆盖，且针对环境变化做了适配，review 中提出的问题均已解决。后续可考虑增加内存监控或资源限制来防止潜在 OOM。

功能与动机

在 NPU Nightly 流水线中，一些测试用例因执行失败或已解决的问题而被跳过。这些测试用例需要适配当前环境并恢复到 Nightly 流水线中进行测试。

实现拆解

实现步骤：

1. 移除跳过标记：在多个测试文件的 register_npu_ci 调用中删除 disabled 参数，使测试重新注册到 nightly 套件。受影响文件包括 test_npu_no_chunked_prefill.py（旧）、test_npu_kimi_vl_a3b_instruct.py、test_npu_llama_3_2_11b_vision_instruct.py、test_npu_openai_function_calling.py、test_npu_deepep_low_latency_qwen3_next.py 等。
2. 重写无分块预填充测试：删除旧文件 test/registered/ascend/basic_function/parameter/test_npu_no_chunked_prefill.py，在 memory_and_scheduling 目录下新建同名文件。新测试类 TestNoChunkedPrefill 改用 setUpClass 启动独立服务器进程、tearDownClass 清理，并使用 run_eval 进行 MMLU 精度评估（阈值 0.65），同时保留 run_bench_serving 的压力测试。
3. 统一 VLM 测试基类：将 TestKimiVLA3BInstruct 从继承 GSM8KAscendMixin 和 CustomTestCase 改为继承共享基类 TestVLMModels，评估数据集从 GSM8K 切换为 MMMU，阈值设为 0.2。移除了不再需要的 other_args（如 --trust-remote-code、--tp-size 等），因为基类已处理。类似地，TestLlama3211BVisionInstruct 也改为从 test_ascend_utils 导入模型路径常量并添加文档字符串。
4. 调整 DeepEP 低延迟测试参数：在 test_npu_deepep_low_latency_qwen3_next.py 中，启用 CUDA Graph 批量大小（--cuda-graph-bs 2 4 6 8）替代 --disable-cuda-graph；设置 DEEPEP_NORMAL_LONG_SEQ_PER_ROUND_TOKENS=3000、DEEPEP_NORMAL_LONG_SEQ_ROUND=10、SGLANG_DEEPEP_BF16_DISPATCH=1 等环境变量以适配 NPU BF16 要求；增大 SGLANG_DEEPEP_NUM_MAX_DISPATCH_TO

KENS_PER_RANK 到 160。

5. 改进函数调用测试：在 test_npu_openai_function_calling.py 中，test_function_call_required 放宽断言：不再要求特定函数名 get_weather 和参数值，仅验证 tool_calls 存在且参数为合法 JSON。添加 strict: true 到工具定义中。

这些改动使之前因跳过而缺失的测试覆盖重新生效。

关键文件：

- test/registered/ascend/basic_function/memory_and_scheduling/test_npu_no_chunked_prefill.py（模块 无分块预填充；类别 test；类型 test-coverage；符号 TestNoChunkedPrefill, setUpClass, tearDownClass, test_mmlu）：新增文件，重写了无分块预填充测试，是本次恢复的核心测试之一。
- test/registered/ascend/basic_function/parameter/test_npu_no_chunked_prefill.py（模块 无分块预填充；类别 test；类型 deletion；符号 TestNoChunkedPrefill, test_no_chunked_prefill, test_no_chunked_prefill_without_radix_cache）：被删除的旧文件，原包含跳过的无分块预填充测试，被新文件替代。
- test/registered/ascend/vlm_models/test_npu_kimi_vl_a3b_instruct.py（模块 Kimi-VL-A3B；类别 test；类型 test-coverage；符号 TestKimiVLA3BInstruct, test_vlm_mmmu_benchmark）：修改了 Kimi-VL-A3B 测试，从 GSM8K 切换为 MMMU 精度评估，并简化配置。
- test/registered/ascend/vlm_models/test_npu_llama_3_2_11b_vision_instruct.py（模块 Llama-3.2-Vision；类别 test；类型 test-coverage）：修改了 Llama-3.2-11B-Vision 测试，使用常量路径并添加文档字符串。
- test/registered/ascend/interface/test_npu_openai_function_calling.py（模块 函数调用；类别 test；类型 test-coverage）：修改了函数调用测试，放宽断言并添加 strict 选项。
- test/registered/ascend/basic_function/parallel_strategy/expert_parallelism/test_npu_deepep_low_latency_qwen3_next.py（模块 DeepEP 低延迟；类别 test；类型 test-coverage）：调整了 DeepEP 低延迟 MoE 测试的环境变量和启动参数。

关键符号：TestNoChunkedPrefill.setUpClass, TestNoChunkedPrefill.tearDownClass, TestNoChunkedPrefill.test_mmlu, TestNoChunkedPrefill.test_no_chunked_prefill_without_radix_cache, TestKimiVLA3BInstruct.test_vlm_mmmu_benchmark, TestLlama3211BVisionInstruct.test_vlm_mmmu_benchmark

评论区精华

Reviewer AndyLi429 提出了四个问题：

- tearDown 缺失：新增的 test_npu_no_chunked_prefill.py 没有进程清理代码。作者在后续提交中添加了 tearDownClass。
- 测试删除疑问：怀疑误删除 test_npu_no_overlap_scheduler.py。作者解释为误操作，已恢复。
- 参数省略：Kimi-VL-A3B 测试移除了 other_args。作者说明因为测试目标从 GSM8K 精度改为 MMMU，使用了基类 TestVLMModels，无需再次指定。

- OOM 风险：担忧 setUpClass 启动服务器后 run_bench_serving 可能内存溢出。作者确认实际测试未触发 OOM。 所有问题均已解决。
- 新增测试缺少 tearDown (correctness): 作者在后续提交中添加了 tearDownClass 清理进程。
- 测试文件删除疑问 (question): 作者解释为误操作，已恢复该文件。
- Kimi-VL-A3B 测试参数省略 (design): 作者说明因为测试从 GSM8K 切换为 MMMU，使用了共享基类 TestVLMModels，无需额外参数。
- setUpClass 与 run_bench_serving 的 OOM 风险 (performance): 作者确认实际测试中未遇到 OOM。

风险与影响

- 风险：主要风险包括：
  1. 无分块预填充测试在同一进程中先后启动服务器和运行 benchmark，若内存不足可能 OOM，但作者验证未发生。
  2. DeepEP 环境变量修改可能影响其他共用 NPU 节点的测试，需要监控稳定性。
  3. 恢复的测试可能会暴露之前隐藏的 flaky 问题，尤其是函数调用测试放宽断言后可能覆盖偏差。 总体风险可控。
- 影响：影响仅限于 NPU Nightly CI 流水线：增加了约 5 个测试用例的回归覆盖，覆盖无分块预填充、多模态 VLM 精度、DeepEP MoE 低延迟部署、OpenAI 函数调用等功能。对 CUDA 或其他硬件后端无影响。团队需在合并后观察 nightly 流水线是否持续稳定。
- 风险标记：OOM 风险，环境依赖，测试 flakiness

关联脉络

- 暂无明显关联 PR