

PR #47231 完整报告

vllm-project/vllm

[XPU][CI] Add tests/v1/e2e/general/test_correctness_sliding_window.py in Intel GPU CI

合并时间: 2026-07-14 16:50

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47231>

执行摘要

- 一句话: Intel GPU CI 集成滑窗测试
- 推荐动作: 该 PR 属于常规 CI 维护, 值得快速合并。建议关注跳过的用例 `test_sliding_window_retrieval[True-1-5-google/gemma-3-1b-it]` 在未来是否得到修复并重新启用。

功能与动机

将滑动窗口注意力测试纳入 Intel GPU CI, 以覆盖该功能的回归测试。PR body 未提供详细动机, 但从变更内容推断是为填补 CI 中滑动窗口功能的测试空白。

实现拆解

1. 修改 CI 配置文件 `.buildkite/intel_jobs/test-intel.yaml`: 在 XPU V1 test 步骤的命令列表中追加一行, 新增 `pytest -v -s v1/e2e/general/test_correctness_sliding_window.py --deselect="tests/v1/e2e/general/test_correctness_sliding_window.py::test_sliding_window_retrieval[True-1-5-google/gemma-3-1b-it]"`。同时, Issue 评论建议将超时时间从 30 分钟延长到 60 分钟, 但最终 PR 将 `timeout_in_minutes` 保持为 70 (已足够)。
2. 修改 CI 脚本 `.buildkite/scripts/hardware_ci/run-intel-ci-test.sh`: 在 `test_suite` 的 case 分支中新增同样的 `pytest` 命令, 确保脚本执行时包含该测试。
3. 跳过已知失败用例: 通过 `--deselect` 跳过 `google/gemma-3-1b-it` 模型的一个特定参数组合, 该用例在 Intel GPU 上已知会失败, 防止 CI 阻塞。

关键文件:

- `.buildkite/intel_jobs/test-intel.yaml` (模块 CI 配置; 类别 config; 类型 configuration): CI 配置文件, 新增滑动窗口测试命令, 是 PR 的核心变更文件。
- `.buildkite/scripts/hardware_ci/run-intel-ci-test.sh` (模块 CI 脚本; 类别 infra; 类型 infrastructure): CI 执行脚本, 同步新增滑动窗口测试命令, 确保脚本方式运行时也包含该测试。

关键符号: 未识别

关键源码片段

`.buildkite/intel_jobs/test-intel.yaml`

CI 配置文件，新增滑动窗口测试命令，是 PR 的核心变更文件。

```
# .buildkite/intel_jobs/test-intel.yaml (head 版本，第 92-104 行)
commands:
  - >-
    bash .buildkite/scripts/hardware_ci/run-intel-test.sh
    'cd tests &&
    pytest -v -s v1/core --ignore=v1/core/test_reset_prefix_cache_e2e.py --ignore=v1/core/test_scheduler_e2e.py &&
    pytest -v -s v1/engine --ignore=v1/engine/test_output_processor.py &&
    pytest -v -s v1/sample --ignore=v1/sample/test_logprobs.py --ignore=v1/sample/test_logprobs_e2e.py -k "not test_topk_only and not test_topp_only and not test_topk_and_topp" &&
    pytest -v -s v1/worker --ignore=v1/worker/test_gpu_model_runner.py --ignore=v1/worker/test_worker_memory_snapshot.py &&
    pytest -v -s v1/structured_output &&
    pytest -v -s v1/test_serial_utils.py &&
    # 新增滑动窗口测试，并跳过已知失败用例 (gemma-3-1b-it 特定参数)
    pytest -v -s v1/e2e/general/test_correctness_sliding_window.py --deselect="tests/v1/e2e/general/test_correctness_sliding_window.py::test_sliding_window_retrieval[True-1-5-google/gemma-3-1b-it]" &&
    pytest -v -s v1/spec_decode --ignore=v1/spec_decode/test_max_len.py --ignore=v1/spec_decode/test_speculators_eagle3.py --ignore=v1/spec_decode/test_acceptance_length.py --ignore=v1/spec_decode/test_speculators_correctness.py &&
    pytest -v -s v1/kv_connector/unit --ignore=v1/kv_connector/unit/test_multi_connector.py --ignore=v1/kv_connector/unit/test_example_connector.py --ignore=v1/kv_connector/unit/test_lmcache_integration.py --ignore=v1/kv_connector/unit/test_hf3fs_client.py --ignore=v1/kv_connector/unit/test_hf3fs_connector.py --ignore=v1/kv_connector/unit/test_hf3fs_metadata_server.py --ignore=v1/kv_connector/unit/test_offloading_connector.py'
```

[.buildkite/scripts/hardware_ci/run-intel-ci-test.sh](#)

CI 执行脚本，同步新增滑动窗口测试命令，确保脚本方式运行时也包含该测试。

```
# .buildkite/scripts/hardware_ci/run-intel-ci-test.sh (head 版本，第 35-44 行)
case "${test_suite}" in
  ...
  pytest -v -s v1/structured_output
  pytest -v -s v1/test_serial_utils.py
  # 新增滑动窗口测试，跳过已知失败用例
  pytest -v -s v1/e2e/general/test_correctness_sliding_window.py --deselect="tests/v1/e2e/general/test_correctness_sliding_window.py::test_sliding_window_retrieval[True-1-5-google/gemma-3-1b-it]"
  pytest -v -s v1/spec_decode --ignore=v1/spec_decode/test_max_len.py ...
  ;;
esac
```

评论区精华

没有 review 评论讨论。Issue 评论中 [wendyliu235](#) 建议将超时时间从 30 分钟延长至 60 分钟，但最终调整为 70 分钟（实际原值即为 70），说明该建议未被直接采纳。

- 超时时间调整 (other): 最终 PR 未修改超时时间（实际保持 70 分钟），说明原有超时已足够，或无需调整。

风险与影响

- 风险：风险极低。修改仅涉及 CI 配置和脚本，无源码逻辑变更。但需注意：新添加的滑动窗口测试可能因为模型下载、GPU 资源或环境差异而不稳定，跳过用例可能导致对应场景在 Intel GPU 上未被完全验证。
- 影响：影响范围极小，仅 Intel GPU CI 的 V1 测试步骤。新增的滑动窗口测试将增加 CI 运行时间约几分钟（取决于测试数量和模型加载时间），但跳过一个用例减少了部分影响。开发者需确保跳过的用例在 Intel GPU 上确实已知失败，并计划修复。
- 风险标记：跳过已知失败用例，CI 时间增加

关联脉络

- 暂无明显关联 PR