PR #26820 完整报告

sgl-project/sglang

Add a sliding-window-attention divergence reporter for the KV-canary

合并时间：2026-05-31 09:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26820

执行摘要

- 一句话：添加 SWA divergence reporter 用于 KV-canary 可观测性
- 推荐动作：值得精读，特别是：
  - DelayedDeviceHostHandler 在设备侧与主机侧异步协作的模式
  - SwaDivergenceLog 的可解析日志格式设计
  - 测试中 assert_swa_divergence_observed 的多信号断言策略 该 PR 展示了如何为关键内部组件添加轻量可观测性基础设施，代码结构清晰，适合作为同类功能的参考。

功能与动机

PR body 指出："Add an opt-in observability reporter for sliding-window-attention (SWA) pools that tracks, per forward, how the SWA verify counts diverge from the FULL pool and how many slots the SWA allocator has remapped, then emits a parseable..."。该功能旨在提供更细粒度的 SWA 行为可见性，帮助调试和验证 SWA 路径是否正确执行。

实现拆解

1. 核心 reporter 实现：在 python/sglang/srt/kv_canary/runner/swa_divergence.py 中新增 SwaDivergenceReporter 类和 SwaDivergenceLog 数据类。SwaDivergenceReporter 在设备上维护一个 shape[2] 的张量（verify_total_count_device）分别累积 FULL 和 SWA 池的验证条目数。每次 forward 通过 observe_after_invoke_plan 将对应池的验证计数累加到该张量。step() 递增前向计数器，并通过 DelayedDeviceHostHandler 异步安排：每 interval 步执行 _compute_on_device 收集 divergence 指标，然后通过 _postprocess_on_host 在主机侧格式化并输出结构化日志。
2. SwaDivergenceLog 数据格式：可冻结数据类包含 forward_ct、verify_full、verify_swa、swa_full_idx_divergence、swa_out_of_window_tokens。format() 方法输出 kv_canary_swa_divergence=... 前缀的日志行，并提供 parse() 和 find_last() 静态方法供测试断言解析。
3. 门控与环境变量：在 python/sglang/srt/environ.py 中添加 SGLANG_KV_CANARY_SWA_DIVERGENCE_STATS_INTERVAL 环境变量（默认 0，标识关闭）。CanaryManager.__init__ 中根据该值条件创建 SwaDivergenceReporter 实例，并传入 SWA 分配器和 req_to_token_pool 引用。

4. 与 SingleForwardManager 的集成：在 single_forward_manager/manager.py 的 pre_ops_maybe_inside_graph 中，每处理一个 buffer group 后调用 reporter.observe_after_invoke_plan；在 PostOpsInsideGraphOutputBuffer 中新增 swa_verify_total_count 字段以便在图中传递设备张量引用。CanaryManager._post_ops_outside_graph 中每步调用 reporter.step。
5. 测试基础设施：在 e2e_base.py 的 CanaryE2EBase 中新增 assert_swa_divergence_observed 方法，检查三个信号（swa_out_of_window_tokens>=1、swa_full_idx_divergence>=1、verify_swa<verify_full）以验证 SWA 路径被真实触发。当 model_mode 为 swa 时，setUpClass 自动设置环境变量间隔为 20。
6. 单元测试：新增 test_self_unit_runner_swa_divergence.py 测试 SwaDivergenceReporter 的日志发射、异步 D2H 传递、以及从日志行解析 SwaDivergenceLog。新增 test_self_unit_e2e_base.py 测试 assert_swa_divergence_observed 的各种条件（通过、阈值比较、重试、无日志等）。已有 SWA 端到端测试中也加入了 maybe_assert_swa_divergence_observed 调用。

关键文件：

- python/sglang/srt/kv_canary/runner/swa_divergence.py（模块 SWA 监测器；类别 source；类型 core-logic；符号 SwaDivergenceReporter, init, observe_after_invoke_plan, step）：核心新增文件，实现了 SwaDivergenceReporter 和 SwaDivergenceLog 类，负责设备端累积、异步 D2H、日志格式化与解析。
- python/sglang/srt/kv_canary/runner/canary_manager.py（模块 KV 监测器；类别 source；类型 dependency-wiring；符号 CanaryManager.init, CanaryManager._post_ops_outside_graph）：作为 reporter 的集成入口，新增 env 门控创建 reporter、在每次 step 中调用 reporter.step、并在 SingleForwardManager 初始化时传递 reporter。
- python/sglang/test/kv_canary/e2e_base.py（模块 测试基类；类别 test；类型 test-coverage；符号 CanaryE2EBase.assert_swa_divergence_observed, CanaryE2EBase.maybe_assert_swa_divergence_observed）：在测试基类中新增 assert_swa_divergence_observed 方法，供端到端测试验证 SWA 路径是否真实执行，提高测试可靠性。

关键符号：SwaDivergenceReporter.init, SwaDivergenceReporter.observe_after_invoke_plan, SwaDivergenceReporter.step, SwaDivergenceReporter._compute_on_device, SwaDivergenceReporter._postprocess_on_host, SwaDivergenceLog.format, SwaDivergenceLog.parse, SwaDivergenceLog.find_last, CanaryE2EBase.assert_swa_divergence_observed

关键源码片段

python/sglang/srt/kv_canary/runner/canary_manager.py

作为 reporter 的集成入口，新增 env 门控创建 reporter、在每次 step 中调用 reporter.step、并在 SingleForwardManager 初始化时传递 reporter。

```
# 在 CanaryManager.__init__ 中添加 reporter 创建（条件门控）
swa_divergence_interval = (
    envs.SGLANG_KV_CANARY_SWA_DIVERGENCE_STATS_INTERVAL.get()
)
if swa_divergence_interval > 0:
    self._swa_divergence_report: Optional[SwaDivergenceReporter] = (
        SwaDivergenceReporter(
            device=device,
            d2h_stream=self._d2h_stream,
            interval=swa_divergence_interval,
            swa_allocator=self._swa_allocator,
            req_to_token_pool=self._req_to_token_pool,
        )
    )
else:
    self._swa_divergence_report = None

# 在 CanaryManager._post_ops_outside_graph 中每步驱动 reporter
if self._swa_divergence_report is not None:
    self._swa_divergence_report.step(
        outer_step_counter=self._outer_step_counter,
        maybe_inaccurate_forward_batch=maybe_inaccurate_forward_batch,
    )
```

评论区精华

该 PR 没有外部 review 评论（review_comments_count=0），由作者直接合入。无公开讨论。

- 暂无高价值评论线程

风险与影响

- 风险：这是一个纯观察者功能，不会修改任何 forward 逻辑或验证结果。主要风险包括：
  - 环境变量误用：若意外设置间隔为正数，会增加 D2H 传输和日志量，但可通过默认 0 关闭。
  - device 张量资源：verify_total_count_device 是一个简单的 int32[2] 张量，开销极低。
  - 与 future map 的交互：DelayedDeviceHostHandler 已在其他 reporter 中使用，机制成熟。
  - 测试覆盖：单元测试覆盖了基本路径，但未测试 CUDA graph 捕获状态下的行为。
  - 影响：影响范围：仅限于 KV-canary 功能的可观测性扩增，不影响正常 serving。影响程度：低。默认关闭，不影响任何现有路径。对开发 / 测试的影响：方便 SWA 相关测试编写和调试，提供可读的 divergence 日志。性能影响：开启后每 interval 步增加一次 D2H 传输和日志输出，资源消耗可忽略。
- 风险标记：暂无

关联脉络

- PR #26821 Add periodic KV-canary stats logging and kernel-run-counter health check: 同系列可观测性增强 PR，增加了常规统计日志和健康检查，与 SWA divergence reporter 共享 DelayedDeviceHostHandler 模式和门控方式。