PR #26590 完整报告

sgl-project/sglang

[BugFix] preserve cached token details in multi-tokenizer output

合并时间：2026-05-29 04:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26590

执行摘要

- 一句话：修复多 tokenizer 路径丢失缓存详情
- 推荐动作：值得精读：变更虽小，但展示了多 tokenizer 路径中字段转发的模式，是维护 metrics 一致性的关键修复。可关注同类字段是否还有遗漏。

功能与动机

修复多 tokenizer 场景下 cached_tokens_details 丢失问题，使 HiCache 逐级缓存指标（device/host/storage_）正确上报。PR body 描述了启动参数和 curl 验证步骤，并指出目前 metrics 仅输出 cache_source="total" 而丢失了预期分层。

实现拆解

1. 源码修改：在 python/sglang/srt/managers/multi_tokenizer_mixin.py 的 _handle_output_by_index 函数中，针对 BatchStrOutput 分支（第 231-248 行），在 cached_tokens 字段之后新增 cached_tokens_details 字段转发，使用现有的 _extract_field_by_index 工具函数按索引提取。
2. 测试新增：新建 test/registered/unit/managers/test_multi_tokenizer_mixin.py 文件，导入 BatchStrOutput 和 _handle_output_by_index。定义 _make_batch_str_output 辅助函数构造包含 cached_tokens_details 的批输出，编写 test_batch_str_output_preserves_cached_tokens_details 测试用例，验证索引为 1 的请求能正确获得 cached_tokens_details=[{"device":1,"host":3}]。
3. CI 注册：测试类通过 register_cpu_ci 注册为 CPU 单元测试，指定预估时长 5 秒，归入 base-a-test-cpu 套件。

关键文件：

- python/sglang/srt/managers/multi_tokenizer_mixin.py（模块 Token 化；类别 source；类型 core-logic；符号 _handle_output_by_index）：核心修复文件，在 BatchStrOutput 分支添加 cached_tokens_details 转发。
- test/registered/unit/managers/test_multi_tokenizer_mixin.py（模块 测试；类别 test；类型 test-coverage；符号 _make_batch_str_output, TestMultiTokenizerMixin, test_batch_str_output_preserves_cached_tokens_details）：新增回归测试，验证 BatchStrOutput 拆分后 cached_tokens_details 正确保留。

关键符号：_handle_output_by_index

关键源码片段

python/sglang/srt/managers/multi_tokenizer_mixin.py

核心修复文件，在 BatchStrOutput 分支添加 cached_tokens_details 转发。

```
# 文件：python/sglang/srt/managers/multi_tokenizer_mixin.py
# 函数 _handle_output_by_index 中 BatchStrOutput 分支（第 231 行起）
# 批量输出按索引 i 拆分为单个请求输出
elif isinstance(output, BatchStrOutput):
    new_output = BatchStrOutput(
        rids=[output.rids[i]],
        spec_verify_ct=_extract_field_by_index(output, "spec_verify_ct", i),
        # ... 其他字段省略 ...
        cached_tokens=_extract_field_by_index(output, "cached_tokens", i),
        # [ 修复 ] 新增 cached_tokens_details 转发，确保 HiCache metrics 能按来源划分
        cached_tokens_details=_extract_field_by_index(
            output, "cached_tokens_details", i
        ),
        input_token_logprobs_val=_extract_field_by_index(
            output, "input_token_logprobs_val", i, check_length=False
        ),
        # ... 后续字段省略 ...
    )
```

test/registered/unit/managers/test_multi_tokenizer_mixin.py

新增回归测试，验证 BatchStrOutput 拆分后 cached_tokens_details 正确保留。

```
# 文件：test/registered/unit/managers/test_multi_tokenizer_mixin.py
import unittest
from sglang.test.ci.ci_register import register_cpu_ci
from sglang.test.test_utils import maybe_stub_sgl_kernel

maybe_stub_sgl_kernel()

from sglang.srt.managers.io_struct import BatchStrOutput
from sglang.srt.managers.multi_tokenizer_mixin import _handle_output_by_index

register_cpu_ci(est_time=5, suite="base-a-test-cpu")

# 构造包含 cached_tokens_details 的批输出，用于测试拆分后字段保留
def _make_batch_str_output() -> BatchStrOutput:
    return BatchStrOutput(
        rids=["rid-0", "rid-1"],
        # ... 其他字段，重点关注 cached_tokens_details
        cached_tokens=[3, 4],
        cached_tokens_details=[
            {"device": 3, "host": 0},
            {"device": 1, "host": 3},
        ],
```

```
        # ... 其余省略
    )

class TestMultiTokenizerMixin(unittest.TestCase):
    def test_batch_str_output_preserves_cached_tokens_details(self):
        output = _make_batch_str_output()
        # 取索引 1 的请求，验证其 cached_tokens_details 正确
        single_output = _handle_output_by_index(output, 1)
        self.assertEqual(single_output.rids, ["rid-1"])
        self.assertEqual(single_output.cached_tokens, [4])
        self.assertEqual(
            single_output.cached_tokens_details,
            [{"device": 1, "host": 3}],
        )

if __name__ == "__main__":
    unittest.main()
```

评论区精华

该 PR 讨论较少，仅有的评论为 bot 配额提醒和测试触发命令。评审人 hnyls2002 直接批准，无反对意见或设计争议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅添加 3 行转发逻辑（使用现有工具函数），不改变其他字段行为。新增测试覆盖拆分正确性。潜在风险：若 cached_tokens_details 在批输出中为 None（非多 tokenizer 场景），_extract_field_by_index 应能正确处理（视其实现而定），但现有测试未覆盖该 case；不过该字段通常由调度器填充，为空时可视为退化情况。
- 影响：影响范围：中。仅影响启用 HiCache、metrics 且 --tokenizer-worker-num >1 的用户。修复后这些用户的 metrics 能正确按缓存来源（device/host/storage）区分，提升可观测性精确性。对其他用户无影响。
- 风险标记：低风险

关联脉络

- PR #26302 [UnifiedTree] gate load back pre-evict on full-attn availability only: 同为 HiCache 缓存管理相关，涉及缓存指标的正确性。