

PR #29882 完整报告

sgl-project/sglang

fix: populate batch req rids and per-request http_worker_ipc for mult...

合并时间: 2026-07-05 03:09

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29882>

执行摘要

- 一句话: 修复 multi-tokenizer 模式下 batch input_ids 路由 hang 住
- 推荐动作: 该 PR 修复了一个重要回归, 且实现简洁清晰, 值得所有涉及 IPC 路由的开发者阅读。特别是 stamp_http_worker_ipc 的设计和 schema 注释的澄清, 展示了 batch 与单请求路由的差异。

功能与动机

Issue #29878 报告 multi-tokenizer 模式启动错误, 请求 hang 住。该 bug 由 PR #29214 引入, 该 PR 添加了 multi-tokenizer 支持, 但 batch tokenized 路径遗漏了 routing 信息的填充, 导致 scheduler 无法将结果返回给正确的 tokenizer worker。

实现拆解

1. 在 python/sglang/srt/managers/tokenizer_manager.py 的 stamp_http_worker_ipc 函数中增加分支: 当对象是 BatchTokenizedGenerateReqInput 或 BatchTokenizedEmbeddingReqInput 时, 遍历其 batch 列表, 为每个子请求设置 http_worker_ipc。这是因为 scheduler 会解包 batch 并读取每个子请求的 http_worker_ipc 来路由响应, 而不是使用父对象的 http_worker_ipcs 列表。
2. 在 python/sglang/srt/managers/io_struct.py 的 BaseBatchReq 类中补充注释, 说明 http_worker_ipcs 字段仅用于 scheduler 输出等并行数组消息, 而 tokenized 输入批量的路由信息存储在 batch[i].http_worker_ipc。同时在 BatchTokenizedGenerateReqInput 和 BatchTokenizedEmbeddingReqInput 类中增加注释, 明确路由方式。
3. 在 test/registered/tokenizer/test_multi_tokenizer.py 中添加 test_batch_input_ids_routing 测试用例, 模拟发送一批 pre-tokenized input_ids (无文本), 验证所有请求都能正常返回响应, 而不会 hang 住。同时调整了 CI 测试的预估时间。

关键文件:

- python/sglang/srt/managers/tokenizer_manager.py (模块 IPC 路由; 类别 source; 类型 core-logic; 符号 stamp_http_worker_ipc): 核心逻辑修复: 在 stamp_http_worker_ipc 中增加对 BatchTokenizedGenerateReqInput 和 BatchTokenizedEmbeddingReqInput 的处理, 为每个子请求设置 http_worker_ipc。
- python/sglang/srt/managers/io_struct.py (模块 数据结构; 类别 source; 类型 documentation; 符号 BaseBatchReq, BatchTokenizedGenerateReqInput,

BatchTokenizedEmbeddingReqInput)：数据结构和注释更新：在 BaseBatchReq 和 BatchTokenized* 类中添加注释，澄清路由方式，避免未来误用。

- test/registered/tokenizer/test_multi_tokenizer.py (模块测试；类别 test；类型 test-coverage；符号 test_batch_input_ids_routing)：添加回归测试 test_batch_input_ids_routing，覆盖 pre-tokenized batch input_ids 路径，确保多 tokenizer 模式下路由正确。

关键符号：stamp_http_worker_ipc, test_batch_input_ids_routing

关键源码片段

python/sglang/srt/managers/tokenizer_manager.py

核心逻辑修复：在 stamp_http_worker_ipc 中增加对 BatchTokenizedGenerateReqInput 和 BatchTokenizedEmbeddingReqInput 的处理，为每个子请求设置 http_worker_ipc。

```
def stamp_http_worker_ipc(obj: Any, ipc_name: str) -> None:
    # 单请求：直接设置 http_worker_ipc
    if isinstance(obj, BaseReq):
        obj.http_worker_ipc = ipc_name
    # 批量 tokenized 输入：scheduler 会解包 batch 并读取每个子请求的
    # http_worker_ipc 来路由响应，因此必须逐个设置
    elif isinstance(
        obj, (BatchTokenizedGenerateReqInput, BatchTokenizedEmbeddingReqInput)
    ):
        for req in obj:
            req.http_worker_ipc = ipc_name
    # 其他批量消息（如 scheduler 输出）：使用 http_worker_ipcs 列表
    elif isinstance(obj, BaseBatchReq):
        obj.http_worker_ipcs = [ipc_name] * len(obj.rids)
```

python/sglang/srt/managers/io_struct.py

数据结构和注释更新：在 BaseBatchReq 和 BatchTokenized* 类中添加注释，澄清路由方式，避免未来误用。

```
class BaseBatchReq(msgspec.Struct, tag=True, kw_only=True, array_like=True):
    # ...
    # Used by batch messages whose items are parallel arrays, such as scheduler
    # outputs. Tokenized input batches store routing on batch[i].http_worker_ipc
    # because the scheduler unpacks them into single-request handlers.
    http_worker_ipcs: Optional[List[Optional[str]]] = None

class BatchTokenizedGenerateReqInput(BaseBatchReq, kw_only=True):
    # The batch of tokenized requests
    # Routing for request i is batch[i].http_worker_ipc, not http_worker_ipcs[i].
    batch: List[TokenizedGenerateReqInput]

class BatchTokenizedEmbeddingReqInput(BaseBatchReq, kw_only=True):
    # Routing for request i is batch[i].http_worker_ipc, not http_worker_ipcs[i].
```

```
batch: List[TokenizedEmbeddingReqInput]
```

test/registered/tokenizer/test_multi_tokenizer.py

添加回归测试 `test_batch_input_ids_routing`，覆盖 pre-tokenized batch input_ids 路径，确保多 tokenizer 模式下路由正确。

```
def test_batch_input_ids_routing(self):
    # Regression guard for sgl-project/sglang#29878 (introduced by #29214).
    #
    # A batch of pre-tokenized `input_ids` (no text / multimodal) is the one
    # case that takes the batch-tokenization path (_send_batch_request ->
    # BatchTokenizedGenerateReqInput). In multi-tokenizer mode this batch
    # must stamp each sub-request's `http_worker_ipc` so the scheduler can
    # route every reply back to its owning tokenizer worker. If it is missing,
    # the requests hang forever.
    #
    # The existing ttft test only sends *text*, so it never exercises this
    # path — this case does, and uses a short timeout so a routing hang
    # fails fast instead of stalling until the server launch timeout.
    batch_input_ids = [
        [1, 2, 3, 4, 5],
        [10, 11, 12, 13, 14],
        [20, 21, 22, 23, 24],
        [30, 31, 32, 33, 34],
    ]
    response = requests.post(
        self.base_url + "/generate",
        json={
            "input_ids": batch_input_ids,
            "sampling_params": {"max_new_tokens": 8, "temperature": 0},
        },
        timeout=60,
    )
    self.assertEqual(response.status_code, 200, response.text)
    results = response.json()
    # Every batched request must get its reply routed back — not hang.
    self.assertEqual(len(results), len(batch_input_ids))
    for result in results:
        self.assertIn("text", result)
```

评论区精华

1. 不要使用 `getattr`: reviewer merrymercy 指出初始实现使用了 `getattr(obj, "batch", None)` 来检查 `batch` 属性，违反了仓库 `.claude/rules/no-getattr-defensive.md` 规范，建议改用 `isinstance` 分支。最终代码直接对 `BatchTokenizedGenerateReqInput` 和 `BatchTokenizedEmbeddingReqInput` 进行 `isinstance` 检查。
2. 避免冗余: merrymercy 质疑为什么需要单独为子请求设置 `http_worker_ipc`，因为 `BaseBatchReq` 已经有 `http_worker_ipcs` 列表。作者解释 scheduler 解包 batch 后按单请

求处理，必须靠子请求自己的 `http_worker_ipc` 字段路由，否则响应无法返回。最终达成一致，并完善了 `io_struct.py` 的注释以澄清两种路由方式的区别。

- 禁止使用 `getattr` 作为防御性检查 (style): 作者将 `getattr` 实现改为明确的 `isinstance` 分支，符合仓库规范。
- 是否冗余：已有 `http_worker_ipcs` 字段 (design): 达成一致，并在 `io_struct.py` 中添加注释澄清两种路由方式，避免冗余设计。最终代码避免了同时设置 `http_worker_ipcs` 和子请求的 `http_worker_ipc`（只设置子请求的）。

风险与影响

- 风险：变更集中在 tokenizer IPC 路由的核心路径，虽然改动量小，但若其他 batch 路径（如未遍历到的子类）未正确处理，可能导致类似 hang 问题。不过当前已通过 `isinstance` 覆盖了所有已知批量 tokenized 输入。测试用例覆盖了关键场景，但未覆盖 embedding 路径 (`BatchTokenizedEmbeddingReqInput`)，存在一定遗漏风险。此外，注释变更无风险。
- 影响：直接影响使用 `--tokenizer-worker-num` 选项启动 multi-tokenizer 模式的用户，修复了请求 hang 问题。其他模式无影响。测试用例的增加提高了 CI 对此类回归的检测能力。
- 风险标记：核心路径变更，多进程通信，回归风险（其他 batch 路径未覆盖）

关联脉络

- PR #29214 [Feature] multi-tokenizer support: 引入该 bug 的 PR，新增了 multi-tokenizer 但未处理 batch tokenized 输入的路由。
- PR #29878 [Bug] multi-tokenizer mode startup error.: 报告该 bug 的 Issue，包含复现命令和截图。
- PR #29928：作者在评论中提到 #29928 可以复现该错误，可能是另一个复现环境或相关修复。