

PR #34769 完整报告

sgl-project/sglang

[AMD][CI] Fix stage-b: AttributeError on multimodal embedding requests

合并时间: 2026-08-15 12:35

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34769>

执行摘要

- 一句话: 修复多模态嵌入请求缺失字段崩溃
- 推荐动作: 该 PR 改动极小但修复了所有多模态 embedding 请求的完全故障, 建议尽快合入。
值得关注的点在于: 公共调用点对请求类型的假设如何导致回归, 以及测试注册平台差异对回归拦截时效的影响。建议后续考虑将覆盖用例加入 NVIDIA PR CI, 或在类型设计上为 EmbeddingReqInput 提供一致接口。

功能与动机

PR #34398 添加了 `_normalize_mm_content_hashes`, 但该函数注解为 `obj: GenerateReqInput`, 其调用点 `_tokenize_one_request` 位于 `if contains_mm_input`: 分支下, 该分支同时被 `GenerateReqInput` 与 `EmbeddingReqInput` 两类请求触达。由于 `mm_content_hashes` 仅在 `GenerateReqInput` 上声明, 多模态 embedding 请求必然触发 `AttributeError: 'EmbeddingReqInput' object has no attribute 'mm_content_hashes'`。该问题被 AMD ROCm 7.2 stage-b CI 的 `test_embedding` 测试捕获 (修复前 2/12 通过)。

实现拆解

1. 问题定位: 在 `python/sglang/srt/managers/tokenizer_manager.py` 的 `_tokenize_one_request` 方法中, 当 `contains_mm_input` 为真时, 会先调用 `_validate_mm_limits(obj)` 再调用 `_normalize_mm_content_hashes(obj)`。后者仅接受 `GenerateReqInput`, 而 `EmbeddingReqInput` 没有 `mm_content_hashes` 字段, 因此每次多模态 embedding 请求都会崩溃。
2. 修复方案: 将 `_normalize_mm_content_hashes(obj)` 调用包在 `isinstance(obj, GenerateReqInput)` 检查中, 确保只有生成请求走内容寻址预处理, 嵌入请求恢复为 #34398 之前的行为。
3. 验证与配套: 未新增测试, 因为 `test/registered/cuda_graph/piecewise/test_piecewise_cuda_graph_support_1_gpu.py` 中的 `TestPiecewiseCudaGraphQwen25VLEmbedding::test_embedding` 已精确覆盖该场景; 在 AMD stage-b shard 上修复后 12/12 通过。无文档与基准变更。

关键文件:

- `python/sglang/srt/managers/tokenizer_manager.py` (模块 请求分词; 类别 `source`; 类型 `core-logic`; 符号 `_tokenize_one_request`): 唯一改动文件, 修复多模态 embedding 请

求在 `_tokenize_one_request` 中因缺少 `mm_content_hashes` 字段而崩溃的问题。

关键符号: `_tokenize_one_request`

关键源码片段

[python/sglang/srt/managers/tokenizer_manager.py](#)

唯一改动文件，修复多模态 embedding 请求在 `_tokenize_one_request` 中因缺少 `mm_content_hashes` 字段而崩溃的问题。

```
# 多模态输入存在时的预处理分支
if contains_mm_input:
    # 校验多模态输入是否超出限制，对生成与嵌入两类请求都适用
    self._validate_mm_limits(obj)

    # mm_content_hashes 是 GenerateReqInput 的字段，EmbeddingReqInput 没有该属性。
    # 仅在生成请求上执行内容寻址哈希归一化，避免多模态嵌入请求
    # 因 AttributeError 崩溃（回归自 #34398）。
    if isinstance(obj, GenerateReqInput):
        self._normalize_mm_content_hashes(obj)
```

评论区精华

mickqian 直接批准 (APPROVED)，无 review 评论。主要讨论集中在 PR 描述与 CI 评论中：作者说明不新增测试的原因——现有 `test_embedding` 已覆盖；同时指出该测试在 AMD 上是 PR 阻塞项，在 NVIDIA 上仅 nightly，导致 #34398 的回归未被及时察觉。amd-bot 确认修复后目标测试通过，其余 CI 失败均与本次改动无关。

- 测试覆盖策略 (testing): 接受不新增测试；作者提出是否应在 NVIDIA PR CI 上也门禁该用例的问题。
- CI 验证结论 (ci): 合并判定为安全，作者请求合并。

风险与影响

- 风险：
 - 类型耦合风险：修复依赖 `GenerateReqInput/EmbeddingReqInput` 的字段差异；若未来 `EmbeddingReqInput` 增加 `mm_content_hashes` 字段，此守卫会使内容寻址功能在嵌入路径失效，需要同步调整。
 - 测试覆盖风险：该回归仅在 AMD stage-b CI 上被拦截，NVIDIA PR CI 不运行对应测试（仅 nightly），类似回归有可能再次溜入。
 - 改动本身风险低：`isinstance` 检查不影响 `generate` 路径，且恢复的是之前一直工作的行为。
 - 影响：影响所有多模态嵌入请求（如 `enable_multimodal=True` 且 `is_embedding=True` 时调用 `engine.encode`），修复前这些请求 100% 崩溃，修复后恢复可用。对生成请求路径无行为变化。对团队而言，消除了 AMD stage-b CI 上持续的红灯，减少排查噪音。
- 风险标记：测试覆盖平台差异，依赖类型字段差异

关联脉络

- 暂无明显关联 PR