

PR #52441 完整报告

vllm-project/vllm

[Bugfix][Multimodal] Keep Gemma 4 video frame counts on CPU

合并时间: 2026-08-16 17:17

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52441>

执行摘要

- 一句话: 修复 Gemma 4 视频帧计数被误搬 GPU 引发的崩溃
- 推荐动作:

功能与动机

PR body 直接指向 Buildkite CI #84014 的引擎崩溃, 并附带了 `dump_input.py` 的完整输入转储。转储显示 Gemma 4 视频请求中 `video_frame_counts` 的字段配置为 `MultiModalBatchedField(keep_on_cpu=False)`, 即该标量 `tensor(16)` 被当作 GPU 输入字段传输; 而同批次的 `video_timestamps`、`video_num_soft_tokens` 均为 `keep_on_cpu=True`。帧计数是 CPU 端按视频拆分扁平帧的元数据, 被搬上 GPU 后破坏了 encoder cudagraph 捕获路径的稳定性假设, 最终触发 `EngineCore` 异常。修复目标就是消除这一字段契约不一致。

实现拆解

1. 修正多模态字段契约 (数据契约)

在 `vllm/model_executor/models/gemma4_mm.py` 的 `_get_mm_fields_config()` 中, 为视频字段 `video_frame_counts` 增加 `keep_on_cpu=True`。该字段是 CPU 端用于 `MultiModalFieldConfig.flat_from_sizes()` 按视频拆分扁平帧的元数据, 此前未标记 `keep_on_cpu` 导致其被当作 GPU 输入字段搬移至设备端, 在 encoder cudagraph 捕获 / 重放路径上与固定 shape 的稳定要求冲突, 触发 Buildkite #84014 的异常转储。修复后该字段与 `video_num_soft_tokens`、`video_timestamps` 两个既有 CPU 元数据字段保持一致, 是纯粹的字段契约对齐。

2. 收紧 gemma4 的 encoder cudagraph 测试预算

在 `tests/models/multimodal/generation/test_vit_cudagraph.py` 中为 `gemma4` 配置新增 `compilation_config_overrides={`

关键文件:

- `vllm/model_executor/models/gemma4_mm.py` (模块 `vllm/model_executor`; 类别 `source`; 类型 `data-contract`): 源码主路径; +1/-1; 以 `data-contract` 为主
- `tests/models/multimodal/generation/test_vit_cudagraph.py` (模块 `vit/cudagraph`; 类别 `test`; 类型 `test-coverage`): 测试配套; 包含 测试覆盖调整、配置键调整; +6/-0

关键符号: 未识别

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR