

PR #45029 完整报告

vllm-project/vllm

Revert "[Bugfix][CI] Gemma3 Transformers multimodal encoder profiling and build prompt-embedding fixtures"

合并时间: 2026-06-10 16:37

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45029>

执行摘要

- 一句话: 回退 Gemma3 多模态编码器 profile 修复
- 推荐动作: 该 PR 维护性质明显, 无需精读。值得关注的是 #39425 的具体修复内容, 以确认 Gemma3 多模态 profile 问题的正确解决方案。

功能与动机

PR #44952 对 Gemma3 的多模态编码器 profile 进行了错误修复, 导致 ROCm 上的 GPU hang 问题和 prompt-embedding fixture 的 VRAM 竞争问题未能正确解决。正确的修复已在 #39425 中完成, 因此需要回退 #44952 中的改动。PR body 明确指出: "Reverts the Transformers backend part of vllm-project/vllm#44952 It was not the correct fix for this problem. <https://github.com/vllm-project/vllm/pull/39425> was."

实现拆解

变更仅涉及一个文件 `vllm/model_executor/models/transformers/multimodal.py`, 回退了 #44952 在该文件中的修改:

1. 删除了新增的辅助方法 `_get_mm_values` 和 `_get_max_encoder_tokens`, 这两个方法是为了在 `get_max_image_tokens` 中针对 Gemma3 模型计算编码器 profile 所需的 patch-token 数量。
2. 修改了 `get_max_image_tokens` 方法, 将返回值从 `self._get_max_encoder_tokens(processor, mm_tokens) or image_tokens` 回退为直接返回 `image_tokens`。
3. 还原了 `get_max_image_size` 方法前的空行, 保持代码格式整齐。

关键文件:

- `vllm/model_executor/models/transformers/multimodal.py` (模块 模型执行器; 类别 source; 类型 data-contract; 符号 `_get_mm_values`, `_get_max_encoder_tokens`): 该文件是 PR 唯一修改的文件, 回退了 #44952 中针对 Gemma3 多模态编码器 profile 的修改, 删除了两个新增方法和相关逻辑。

关键符号: `get_max_image_tokens`, `_get_mm_values`, `_get_max_encoder_tokens`

评论区精华

该 PR 没有 review 评论。唯一的审核来自 AndreasKaratzas，批准了该回退，表明团队一致认为 #44952 的修复不正确。

- 暂无高价值评论线程

风险与影响

- 风险：该回退本身风险较低，因为它只是撤销了一个已知错误的修复，并将代码恢复到之前的状态。但需注意，如果 #44952 修复的部分问题尚未通过 #39425 完全解决，Gemma3 在 ROCm 上的多模态 profile 仍可能存在问题。不过，PR 作者明确指向 #39425 作为正确修复，因此风险可控。
- 影响：影响范围有限，仅涉及 Gemma3 模型在 Transformers 后端上的多模态编码器 profile 逻辑。回退后，`get_max_image_tokens` 不再考虑 Gemma3 的 vision encoder 预池化 patch-token 数量，而是直接使用 `_get_num_multimodal_tokens` 返回的 `num_image_tokens`。这可能导致 Gemma3 的 profile 内存估算更加保守（更小），但不会影响推理正确性。
- 风险标记：暂无

关联脉络

- PR #44952 [Bugfix][CI] Gemma3 Transformers multimodal encoder profiling and build prompt-embedding fixtures: 本 PR 直接回退 #44952 的变更，认为该修复不正确。
- PR #39425 Unknown (referenced in PR body): PR body 指出 #39425 是 Gemma3 多模态 profile 问题的正确修复，本回退依赖该 PR 的后续合并。