PR #39268 完整报告

vllm-project/vllm

[Tests] Add Qwen3-VL multimodal memory leak check

合并时间：2026-04-09 19:54

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39268

执行摘要

本 PR 新增了针对 Qwen3-VL 多模态模型的内存泄漏检测测试，通过重复发送固定请求并监控 GPU 和 CPU 内存增长，旨在早期发现潜在泄漏，作为 issue #16353 多模态性能基准测试计划的一部分，增强 CI 测试覆盖和模型稳定性验证。

功能与动机

此变更源于 issue #16353（“为多模态模型在 CI 中运行性能基准测试”），旨在避免多模态优化（如处理器缓存）导致的回归，如内存泄漏。PR body 明确指出，专注于 Qwen3-VL-4B-Instruct 模型，添加引擎级泄漏检查，而非完整服务路径基准测试。具体地，通过建立内存基线和检查后续轮次增长，提供针对性泄漏检测能力。

实现拆解

核心文件变更

- tests/models/multimodal/generation/test_memory_leak.py：新增测试文件，包含以下关键逻辑：
  - 使用 LLM 初始化模型，设置多模态参数（如 limit_mm_per_prompt）。
  - 定义 _make_messages 和 _build_request_batch 函数构建包含随机文本和图像的请求，避免前缀缓存干扰。
  - 通过 _gpu_used_bytes 获取 GPU 使用内存，_ru_maxrss_bytes 获取 CPU 峰值 RSS。
  - 执行多轮请求（预热轮和测量轮），检查内存增长是否超过阈值（初始设为 0 MiB，但 review 建议调整）。
  - 使用 pytest.mark.core_model 标记，并支持参数化图像输入。
- vllm/utils/mem_constants.py：添加 KB_bytes 和 KiB_bytes 常量定义，为内存单位提供标准基准。
- vllm/utils/mem_utils.py：新增 format_kib 函数，用于格式化内存值为 KiB 单位，增强测试输出可读性。

关键函数

- _make_messages：构建包含随机文本和图像 URL 的聊天消息，避免缓存效应。
- _build_request_batch：生成固定数量的请求批次，用于每轮测试。

- _ru_maxrss_bytes：跨平台获取进程峰值 RSS（Linux 为 KB，macOS 为字节）。
- _gpu_used_bytes：同步设备并计算 GPU 已使用内存。
- format_kib：新增工具函数，格式化字节值为 KiB 字符串。

评论区精华

review 中，gemini-code-assist[bot] 提出两个高优先级讨论：

> “内存增长阈值（256 MiB for GPU 和 128 MiB for CPU）过高，对于固定输入的 4B 模型，内存占用应极稳定，高阈值可能掩盖显著泄漏（如每请求 10 MiB），建议降低到更敏感值如 32 MiB 或 64 MiB。”
>
> “使用 resource.getrusage 的峰值 RSS 可能隐藏初始化后的泄漏，因为峰值通常在模型加载时达到，建议跟踪当前 RSS（例如使用 psutil）以有效检测测量阶段的泄漏。”

讨论未显示作者明确回应，但 PR 最终被 DarkLight1337 批准，暗示可能已采纳建议或在后续提交中调整阈值和测量方法。

风险与影响

技术风险

- 阈值设置不当：初始阈值过高可能导致小泄漏漏检，而调整过低可能引入误报，影响测试可靠性。
- 测量方法局限：依赖峰值 RSS 可能无法捕捉模型加载后的泄漏，需结合当前 RSS 跟踪以提高准确性。
- 模型依赖性强：测试仅针对 Qwen3-VL-4B-Instruct，缺乏通用性，可能不覆盖其他多模态模型变体。
- 环境干扰：GPU 内存波动或系统负载可能影响测量结果，导致测试不稳定。

影响范围

- 用户：无直接影响，此为测试代码变更。
- 系统：增强 CI 测试套件，为多模态模型提供自动化内存泄漏检测，有助于早期发现回归问题。
- 团队：提升开发效率，减少手动内存检查负担，并推动多模态性能基准测试的持续集成进程。

关联脉络

本 PR 直接关联 issue #16353，是“多模态模型性能基准测试”功能请求的具体实现之一。从同仓库近期历史 PR 看，多模态相关变更（如 #38388 修复嵌套张量比较、#39307 更新 ColModernVBERT 模型）可能受益于此测试覆盖，形成多模态稳定性验证的演进链条。此外，review 讨论中提及的阈值和测量方法改进，可能为后续类似测试（如其他模型或更全面的基准测试）提供设计参考。