PR #44615 完整报告

vllm-project/vllm

[Bugfix] Fix gemma4 crash on CPU: guard mem_get_info call

合并时间：2026-06-05 20:47

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44615

执行摘要

- 一句话：为 CpuPlatform 补充 mem_get_info，修复 gemma4 CPU 崩溃
- 推荐动作：本 PR 虽然代码量小，但展示了一个很好的设计决策过程：从临时"补丁"到根本修复的转变。对于理解 vLLM 平台抽象层（Platform 类体系）以及如何优雅地处理平台差异，这是一个值得学习的小案例。建议平台层相关开发者阅览。

功能与动机

关联 Issue #44039 报告了 gemma4 在 CPU 上运行时 _process_video_input 无法使用的问题。根本原因是 CPU 平台缺少 mem_get_info 方法，导致多模态编码分块逻辑无条件调用一个实际为 None 的可调用对象而崩溃。该修复是让 gemma4 多模态处理在 CPU 上正常运行的前提步骤。

实现拆解

1. 问题定位：排查发现 current_platform.mem_get_info 在 CPU 平台返回 None（因为 Platform.__getattr__的 fallback 机制），导致 TypeError。
2. 初始修复尝试：贡献者最初在 gemma4_mm.py 中通过 callable() 守卫临时避免崩溃。
3. Reviewer 指导：bigPYJ1151 指出更根本的解决方法是在 CpuPlatform 中实现 mem_get_info 方法，并利用 cpu_resource_utils.get_memory_node_info 获取真实内存数据。
4. 最终实现：在 vllm/platforms/cpu.py 中添加了 @classmethod mem_get_info，返回 (available_memory, total_memory) 元组。
5. 测试清理：应 reviewer 要求，移除了之前为测试新增的测试函数。

关键文件：

- vllm/platforms/cpu.py（模块 平台层；类别 source；类型 core-logic；符号 mem_get_info）：唯一变更文件，为 CpuPlatform 添加 mem_get_info 类方法，解决崩溃根因。

关键符号：CpuPlatform.mem_get_info

关键源码片段

vllm/platforms/cpu.py

唯一变更文件，为 CpuPlatform 添加 mem_get_info 类方法，解决崩溃根因。

```
# vllm/platforms/cpu.py 中 CpuPlatform 类的部分

@classmethod
def get_device_total_memory(cls, device_id: int = 0) -> int:
    # 已有方法，展示了 get_memory_node_info 的使用方式
    meminfo = get_memory_node_info(device_id)
    return meminfo.total_memory

@classmethod
def mem_get_info(cls) -> tuple[int, int]:
    """
    返回 (available_memory, total_memory) ，单位字节。
    该方法与 torch.cuda.mem_get_info 接口保持一致，
    用于满足多模态编码分块时对当前可用内存的查询需求。
    """
    meminfo = get_memory_node_info() # 默认 device_id=0
    return meminfo.available_memory, meminfo.total_memory
```

评论区精华

设计讨论：bigPYJ1151 提出不应只在调用处加保护，而应在 CpuPlatform 中完整实现 mem_get_info。贡献者采纳此建议，彻底修复问题。

测试讨论：bigPYJ1151 认为新增的测试不必要，要求删除。贡献者遵从反馈移除了测试。

- 从调用处保护转向 CpuPlatform 实现 (design): 贡献者 adhithyamulticoreware 接受了建议，修改为在 CpuPlatform 中添加 mem_get_info 方法，调用 get_memory_node_info 实现。
- 移除不必要的测试 (testing): 贡献者移除了测试，并确认完成。

风险与影响

- 风险：主要风险：mem_get_info 实现中调用 get_memory_node_info() 未传递 device_id 参数，而同类的 get_device_total_memory 却传递了 device_id。对于多 CPU 节点环境，当前实现可能始终查询节点 0，无法感知其他节点的内存状态。不过 vLLM 的 CPU 后端目前仅支持单节点，此风险在当前使用场景下较低。

其他风险：缺少针对 mem_get_info 返回值的回归测试，但 reviewer 明确认为不需要测试，因此该风险被接受。

- 影响：直接影响：gemma4 多模态处理（图像 / 视频）现在可在 CPU 平台上正常执行，不会再因 TypeError 崩溃。

对 GPU 等其他平台无任何影响。变更仅涉及 cpu.py 一个文件，且仅新增一个方法，侵入性极小。用户无需调整使用方式即可受益。

- 风险标记：暂无

关联脉络

- 暂无明显关联 PR