PR #37533 完整报告

vllm-project/vllm

[ROCm] fix sleep mode not releasing GPU memory problem on ROCm

合并时间：2026-03-23 21:07

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37533

执行摘要

此 PR 修复了在 ROCm 平台上睡眠模式无法释放 GPU 内存的问题，通过在 unmap_and_release 函数中添加虚拟地址循环操作强制物理内存释放，影响 ROCm 用户的内存管理功能，但可能存在性能优化空间。

功能与动机

在 ROCm 上，hipMemRelease 不会将物理 VRAM 返回空闲池，导致睡眠模式下 GPU 内存保持占用，阻止其他应用使用释放的内存。PR body 明确指出此问题，并提供脚本复现和验证。

实现拆解

仅修改 csrc/cumem_allocator.cpp 文件中的 unmap_and_release 函数，针对 ROCm 路径添加以下代码块：

```
if (first_error == no_error) {
    first_error = cuMemAddressFree(d_mem, size);
    if (first_error == no_error) {
        CUdeviceptr d_mem_new = 0;
        first_error = cuMemAddressReserve(&d_mem_new, size, 0, d_mem, 0);
        // 错误处理逻辑...
    }
}
```

该实现确保释放物理页并保留相同虚拟地址，以兼容后续唤醒操作。

评论区精华

review 中，gemini-code-assist[bot] 指出：

> 此工作区对睡眠路径正确，但由于 unmap_and_release 也被 my_free 调用，在普通释放路径中会导致不必要的内存操作，建议重构以隔离工作区到仅睡眠路径。 tjtanaa 批准 PR 但未采纳此建议，表明设计权衡未完全解决。

风险与影响

- 性能风险：在 my_free 路径中增加额外内存操作，可能影响普通张量释放效率。
- 兼容性风险：仅针对 ROCm 路径，CUDA 路径不受影响，但错误处理依赖特定地址返回。

- 影响范围：使用 ROCm 平台并启用睡眠模式的用户将受益于正确内存释放，提升系统资源利用率。

关联脉络

与此 PR 相关的历史 PR 如 #36100 和 #36505 均涉及 ROCm 平台的错误修复和性能优化，揭示 vLLM 仓库对 ROCm 支持的持续演进，尤其是在内存管理和内核优化方面。