PR #22422 完整报告

sgl-project/sglang

[AMD] Replace triton rotary_emb with aiter rotary_emb for Wan2.2 denoise

合并时间：2026-04-10 09:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22422

执行摘要

本 PR 将 WanVideo 模型在 AMD 平台上的 RoPE（旋转位置编码）实现从 Triton 内核替换为 aiter HIP 内核，通过减少内存分配和形状转换开销，实现内核级约 40% 的性能提升，端到端去噪时间减少约 0.5%。变更已通过准确性测试验证，风险较低，是 AMD 平台多模态模型性能优化的典型案例。

功能与动机

为什么做？原 Triton 实现 rope_cached_thd_positions_2c_fwd 每次调用需分配两个新输出张量，并需进行 reshape/view_as 操作，导致性能开销。aiter HIP 内核 rope_cached_2c_fwd_inplace 能以单内核完成 k 和 v 的 RoPE 计算，且支持原位操作，从而优化内存使用和计算效率。

解决什么问题？提升 WanVideo 模型在 AMD 平台上的推理性能，减少去噪步骤的延迟。PR body 中提供了详细的基准测试数据：内核级时间从 2,922.2 µs 降至 1,334.6 µs，端到端去噪时间从 115.14 秒降至 114.68 秒。

实现拆解

实现涉及两个文件的关键改动：

1. 配置层(python/sglang/multimodal_gen/runtime/models/utils.py): python _use_aiter = get_bool_env_var("SGLANG_USE_AITER") and _is_hip 新增 _use_aiter 标志，基于环境变量和 HIP 平台检测动态启用 aiter 优化。
2. 核心逻辑层(python/sglang/multimodal_gen/runtime/models/dits/wanvideo.py):
   - 在 forward 方法中，当 _use_aiter 为真时，替换原有 _apply_rotary_emb 调用。
   - 关键步骤：python q_sbhd = query.view(num_tokens, 1, query_shape[-2], query_shape[-1]) k_sbhd = key.view(num_tokens, 1, key_shape[-2], key_shape[-1]) rope_cached_2c_fwd_inplace(q_sbhd, k_sbhd, cos_sbhd, sin_sbhd, ...) 将张量重塑为 sbhd 格式后调用 HIP 内核，再恢复原始形状。

评论区精华

Review 中唯一但关键的讨论围绕 导入安全性展开：

> HaiShaw: "Please not to do conditionless aiter import, do within _is_hip."
> Jackycheng0808: "Remove the unconditional aiter import; import it only when use_aiter is enabled..."

核心洞察：初始提交使用 try: 无条件导入 aiter 模块，可能在非 HIP 环境引发错误。通过将导入移至 _use_aiter 条件内，确保平台兼容性。这体现了在平台特异性优化中维护代码健壮性的最佳实践。

风险与影响

风险：

- 兼容性：依赖 aiter 库和 HIP 平台，若环境配置不当可能回退到原有路径，需确保回退逻辑正确。
- 正确性：尽管准确性测试通过（CLIP 均值 0.9899、SSIM 均值 0.8657 等指标达标），但内核替换需确保边界情况处理一致。
- 性能收益：端到端改进仅 0.5%，可能受其他瓶颈限制，实际收益需生产验证。

影响：

- 用户：AMD 平台用户获得轻微性能提升，功能不变。
- 系统：减少内存分配和内核调用，潜在降低系统开销。
- 团队：延续 AMD 优化脉络，与近期 PR（如 #22424 的 LayerNorm 优化）形成技术协同。

关联脉络

本 PR 是 AMD 平台性能优化系列的一部分：

- #22424：使用 aiter CK layernorm2d 减少 NSA 索引器内核启动，与本 PR 同为内核替换优化。
- #22329：为 AMD MORI-EP 新增环境变量，与本 PR 共享通过环境变量控制平台特性的模式。
- #22089：为 Qwen3-ASR 添加流式语音识别，反映多模态领域的并行投入。

演进趋势：sglang 项目正持续深化平台特异性优化（尤其是 AMD），通过内核级替换（Triton→aiter/HIP）提升性能，同时注重通过环境变量和条件导入保持代码可维护性。