PR #21234 完整报告

sgl-project/sglang

[AMD] Support AMD MXFP4 Qwen3.5-397B-A17B model

合并时间：2026-03-30 16:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21234

执行摘要

本 PR 为 AMD GPU 添加了 Qwen3.5 MXFP4 量化模型支持，通过修改模型代码引入融合模块映射，提升了推理性能约 18% 吞吐量和 15% 延迟降低，但准确率轻微下降，需后续优化。

功能与动机

动机是启用 AMD GPU 上的 Qwen3.5 MXFP4 模型推理，以在保持可接受准确率的同时，提升服务性能。引用 PR body："Enable and validate Qwen3.5 MXFP4 model support on AMD GPUs This PR aims to preserve acceptable accuracy while improving serving performance versus the FP8 baseline." 基准测试显示，与 FP8 基线相比，MXFP4 模型在 GSM8K 任务上平均准确率从 0.9495 降至 0.9315，但延迟和吞吐量显著改善。

实现拆解

改动集中在 python/sglang/srt/models/qwen3_5.py 文件，涉及以下关键变更：

- 引入 _is_gfx95_supported() 函数检测 AMD GPU 支持。
- 在 Qwen3_5ForCausalLM 类中添加 packed_modules_mapping 映射，包含 Quark 特有的融合模块名（如 in_proj_qkvz 映射到 ['in_proj_qkv', 'in_proj_z']）。
- 在 Qwen3_5ForConditionalGeneration 和 Qwen3_5MoeForConditionalGeneration 类中继承或设置相同映射，确保多模态和 MoE 变体兼容。
- 代码示例：

评论区精华

- 条件检查争议：BowenBao 质疑："is _is_hip required or can this be relaxed for all quant_config?" hubertlu-tw 初始回应为保持其他代码路径不变，后移除以使映射硬件无关。最终代码移除检查，提升通用性。
- 准确率担忧：HaiShaw 指出："model amd/Qwen3.5-397B-A17B-MXFP4 yields substantial lower accuracy scores"，BowenBao 回应可能是集成问题，在其他框架中恢复率较高。此问题未解决，需跟踪。

风险与影响

- 风险：准确率下降可能影响模型输出质量，需监控和优化；兼容性风险，修改可能意外影响非 AMD 配置，但讨论后已降低；缺少单元测试，PR checklist 中测试项未完成，可能隐藏回归；核心模型路径变更，在 qwen3_5.py 中添加逻辑，需确保在所有场景下正确工作。
- 影响：AMD 用户受益于性能提升，但需权衡准确率牺牲；系统层面优化推理效率，增强 SGLang 对量化模型的支持；团队扩展 AMD 生态，但需处理准确率差距，可能涉及跨框架对齐。

关联脉络

与历史 PR 的关联揭示量化支持和 AMD 优化的演进趋势：

- PR #21448：修复 Qwen3.5 MoE 模型加载问题，修改相同文件，共享模型代码维护上下文。
- PR #14385：实现 MXFP4 Gemm 内核用于 Intel CPU，共享量化技术（MXFP4），反映跨硬件量化支持的一致性努力。
- PR #21315：AMD GPU 的 RoPE 与 KV 缓存融合优化，同属 amd 标签系列，显示 AMD 生态的持续性能优化。