PR #37247 完整报告

vllm-project/vllm

[Model] Implement LoRA support for Qwen3ASRForConditionalGeneration

合并时间：2026-04-10 22:34

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37247

执行摘要

此 PR 为 Qwen3-ASR 模型添加了 LoRA 支持，通过实现接口、调整音频塔层和修复模型加载逻辑，解决了之前 LoRA 适配器无法工作的问题，并更新了相关文档。

功能与动机

动机源于 Issue #37223，用户在使用 LoRA 适配器服务 Qwen3-ASR 模型时遇到错误 'Qwen3ASRForConditionalGeneration does not support LoRA yet.'，此 PR 旨在消除此障碍，使模型能够集成 LoRA 适配器。

实现拆解

关键改动点包括：

- vllm/model_executor/models/qwen3_asr.py：添加 SupportsLoRA 接口和 packed_modules_mapping，并实现 get_num_mm_encoder_tokens() 方法。
- vllm/model_executor/models/qwen3_omni_moe_thinker.py：将音频塔中的 nn.Linear 替换为 ReplicatedLinear，示例如下：python self.conv_out = ReplicatedLinear(conv_out_dim, config.d_model, bias=False, return_bias=False, prefix=f"{prefix}.conv_out")
- vllm/v1/worker/gpu_model_runner.py：修复条件判断，从检查 hasattr(self.model, "get_num_mm_connector_tokens") 改为检查 connector 实际存在。
- docs/models/supported_models.md：更新文档，标记 Qwen3ASRForConditionalGeneration 支持 LoRA。

评论区精华

jeejeelee 在 review 中提出关键问题：

> "QQ: why add this?" 关于 embedding_modules 变量。

作者 petern48 回应解释了音频塔的集成问题，并参考了其他模型实现。讨论聚焦于如何正确处理多模态模型中的 LoRA 支持。

风险与影响

- 技术风险：音频塔线性层替换可能引入性能回归，但使用 ReplicatedLinear 是标准做法；条件判断修复可能影响其他多模态模型，需谨慎验证。
- 影响范围：用户现在可以对该模型使用 LoRA 适配器，增强了模型定制性；系统层面影响多模态处理流程，特别是音频塔的 LoRA 集成。

关联脉络

此 PR 与近期其他模型支持 PR（如 #39290 FireRedLID）相似，属于 vLLM 扩展多模态模型生态的一部分。关联 Issue #37223 直接驱动了此次变更，表明社区需求推动了功能迭代。