PR #50434 完整报告

vllm-project/vllm

[XPU] [BugFix] Add deepseek_v4_fp8 to xpu supported_quantization list

合并时间：2026-07-31 11:12

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50434

执行摘要

- 一句话：XPU 支持 deepseek_v4_fp8 量化格式
- 推荐动作：对于使用 Intel XPU 的开发者，此 PR 值得关注，解决 DeepSeek V4 FP8 模型无法加载的问题。实现简单直接，意义明确。

功能与动机

修复 XPU 平台不支持 deepseek_v4_fp8 量化格式的问题，原因是用户在使用 XPU 设备加载 DeepSeek V4 FP8 模型时，因 supported_quantization 列表中缺少该格式而失败。PR 中引用了相关 PR #44513，说明该格式在其他平台已支持，此处补充 XPU 的适配。

实现拆解

1. 修改 vllm/platforms/xpu.py 中 XPUPlatform 类的 supported_quantization 属性，在 'fp8' 之后新增 'deepseek_v4_fp8' 字符串。
2. 该列表用于平台量化格式的合法性校验，新增后 XPU 平台将接受 deepseek_v4_fp8 量化配置。
3. 无新增测试，改动依赖现有量化路径，属于配置变更。

关键文件：

- vllm/platforms/xpu.py（模块 平台；类别 source；类型 core-logic；符号 XPUPlatform.supported_quantization）：在 XPUPlatform 的 supported_quantization 列表中添加 deepseek_v4_fp8，使 XPU 平台支持该量化格式，是整个 PR 的核心变更。

关键符号：XPUPlatform.supported_quantization

关键源码片段

vllm/platforms/xpu.py

在 XPUPlatform 的 supported_quantization 列表中添加 deepseek_v4_fp8，使 XPU 平台支持该量化格式，是整个 PR 的核心变更。

```
# vllm/platforms/xpu.py
class XPUPlatform(Platform):
    # ...
    # 该列表用于校验用户指定的量化格式是否被当前平台支持
```

```
    supported_quantization: list[str] = [
        "awq",
        "gptq",
        "auto_awq",
        "auto_gptq",
        "inc",
        "fp8",
        "deepseek_v4_fp8", # 新增：允许 XPU 使用 DeepSeek V4 的 FP8 量化
        "mxfp4",
        "mxfp8",
        "fp8_per_tensor",
        "fp8_per_block",
        "online",
        "gpt_oss_mxfp4",
        "modelopt",
        "compressed-tensors",
    ]
```

评论区精华

无实质 review 评论，仅有 Claude bot 的自动提示（因 fork 仓库未启用自动化审查），最终由维护者 jikunshang 和 yma11 批准合并，说明改动简单且可靠。

- 暂无高价值评论线程

风险与影响

- 风险：该改动仅增加配置项，不涉及逻辑变更，风险极低。但需确认 XPU 后端实际实现了 deepseek_v4_fp8 量化所需的算子和内核（如 FP8 权重转换），否则即使配置通过，运行期仍可能失败。建议在 XPU 环境验证模型加载和推理。
- 影响：仅影响 XPU 平台，允许使用 deepseek_v4_fp8 量化格式，扩大 XPU 的模型支持范围。对现有功能和性能无影响，改动极小，风险低。
- 风险标记：缺少测试覆盖，需验证后端实现

关联脉络

- PR #44513 相关的 deepseek_v4_fp8 支持 PR: PR body 中引用了该 PR，说明 deepseek_v4_fp8 的原始支持实现，本 PR 是对其在 XPU 平台的补充。