PR #42163 完整报告

vllm-project/vllm

Document MolmoWeb hf_overrides

合并时间：2026-05-11 14:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42163

执行摘要

- 一句话：更新文档说明 MolmoWeb 模型使用 hf_overrides 配置
- 推荐动作：建议所有使用 MolmoWeb 模型或类似架构变体的用户阅读此文档。对于开发者，这是一个很好的示例，展示了如何在不添加新代码的情况下通过文档和 hf_overrides 机制支持模型变体。

功能与动机

根据 issue #38660 报告，MolmoWeb 模型在使用 Molmo2 架构时因多模态注意机制差异导致 CUDA 断言错误。实际根因是 MolmoWeb 训练时使用因果注意力而非多模态前缀注意。本 PR 通过文档指导用户显式设置 hf_overrides 来覆盖 is_mm_prefix_lm 参数，避免用户自行添加新架构代码。

实现拆解

该 PR 仅修改了一个文档文件 docs/models/supported_models.md：

1. 在 Molmo2 支持模型的列表中增加 allenai/MolmoWeb-4B 和 allenai/MolmoWeb-8B，并附上上标 ^ 引用底部注释。
2. 在文档末尾的注意事项区域添加一条 note，说明使用 MolmoWeb 时需要设置 --hf-overrides 将架构设为 Molmo2ForConditionalGeneration 并关闭 is_mm_prefix_lm。
3. 调整了注释块的顺序以保持字母排序。

没有新增任何代码文件，完全通过文档方式向用户传递用法。

关键文件：

- docs/models/supported_models.md（模块 文档；类别 docs；类型 documentation）：唯一变更文件，更新了模型支持列表并添加了 MolmoWeb 使用说明。

关键符号：未识别

评论区精华

在 review 过程中，维护者 DarkLight1337 建议只做文档变更，而不是添加新的模型包装器 (molmoweb.py)，因为当前 MolmoWeb 检查点仍标记为 Molmo2ForConditionalGeneration 架构，直接在文档中说明 hf_overrides 更符合 vLLM 对其他类似模型的现有做法。最终 PR 合并时只包含了文档变更。

- 是否添加 MolmoWeb 模型包装器 vs 仅文档变更 (design): 采纳仅文档变更方案，不添加新代码。

风险与影响

- 风险：这是一个纯文档变更，不涉及任何代码修改，因此无回归或性能风险。但需要确保用户理解并正确使用 hf_overrides 参数，否则可能导致模型行为不符合预期。
- 影响：对用户：有指导作用，帮助 MolmoWeb 用户正确配置 vLLM。对系统：无影响。对团队：降低了维护成本，无需为每个变体新增代码。
- 风险标记：纯文档变更，无代码风险

关联脉络

- PR #38660 [Bug]: CUDA assert in triton attention for MolmoWeb models: 关联 issue，本 PR 的动机来源于此 bug 报告，最终以文档方式指导用户解决。
- PR #42162 Fix Molmo2 image token metadata: 本 PR 堆叠在此修复之上，先修复了 Molmo2 图像令牌元数据问题，再添加 MolmoWeb 文档。