PR #40818 完整报告

vllm-project/vllm

[Test] Increase qwen2_vl num_logprobs to fix torch 2.12 update

合并时间：2026-04-25 08:59

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40818

执行摘要

- 一句话：增加 Qwen2-VL 测试的 num_logprobs 参数以兼容 torch 2.12
- 推荐动作：建议合并。虽然未深入根因，但作为测试配置调整已足够。后续可跟踪 torch 2.12 相关 issue 以根本解决。

功能与动机

在 torch 2.12 升级过程中，test_multi_image_models[qwen2_vl-test_case43] 测试在 L4 GPU 上失败，表现为 HF 参考与 vLLM 输出在细微处分歧（如 'window' vs 'archway'）。问题根源可能涉及 torch 2.12 下浮点精度或随机采样行为的变化。PR 作者发现提高 num_logprobs 到 10 可修复该问题，且其他模型已采用此值。

实现拆解

1. 在 tests/models/multimodal/generation/test_common.py 的 qwen2_vl 测试配置中新增 num_logprobs=10 参数。
2. 该参数控制测试时对比 logprobs 的 top-k 数量，默认值为 5（未显式指定时），增大该值可降低采样随机性导致的输出偏差。
3. 此变更仅影响测试断言逻辑，不涉及模型推理或训练代码，无需额外依赖或配置。

关键文件：

- tests/models/multimodal/generation/test_common.py（模块 测试；类别 test；类型 test-coverage）：测试配置入口，新增 num_logprobs=10 以稳定 Qwen2-VL 测试结果

关键符号：未识别

关键源码片段

tests/models/multimodal/generation/test_common.py

测试配置入口，新增 num_logprobs=10 以稳定 Qwen2-VL 测试结果

```
# 文件：tests/models/multimodal/generation/test_common.py
# 在 qwen2_vl 的 VLMTestInfo 定义中，添加 num_logprobs=10
# 此参数控制测试时对比 logprobs 的 top-k 数量，默认 5，增大可降低采样随机性
"qwen2_vl": VLMTestInfo(
    models=["Qwen/Qwen2-VL-2B-Instruct"],
```

```
    test_type=(VLMTestType.IMAGE, VLMTestType.MULTI_IMAGE, VLMTestType.VIDEO),
    prompt_formatter=lambda img_prompt: f"<|im_start|>User\n{img_prompt}<|im_end|>\n<|im_
    start|>assistant\n",
    img_idx_to_prompt=lambda idx: "<|vision_start|><|image_pad|><|vision_end|>",
    video_idx_to_prompt=lambda idx: "<|vision_start|><|video_pad|><|vision_end|>",
    multi_image_prompt="Picture 1: <vlm_image>\nPicture 2: <vlm_image>\nDescribe these two
    images with one paragraph respectively.",
    max_model_len=4096,
    max_num_seqs=2,
    num_logprobs=10, # 新增：提高 logprobs 数量以稳定 torch 2.12 下的测试结果
    auto_cls=AutoModelForImageTextToText,
    vllm_output_post_proc=model_utils.qwen2_vllm_to_hf_output,
    image_size_factors=[(0.25,), (0.25, 0.25, 0.25), (0.25, 0.2, 0.15)],
    marks=[pytest.mark.cpu_model],
),
```

评论区精华

无有效 review 讨论，仅包含自动化 bot 确认和 maintainer 的批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅针对特定测试用例的参数调整，不影响生产代码。但需要注意：该修复未解释根本原因（torch 2.12 的底层行为变化），未来 torch 版本升级后可能需重新评估。
- 影响：对用户无影响，对测试稳定性有直接改善：消除 Qwen2-VL 在 L4 GPU 上的偶发测试失败，确保 torch 2.12 升级进程不受阻塞。
- 风险标记：未探明根因，仅测试调整

关联脉络

- PR #40077 [Upgrade] torch 2.12 upgrade tracking: 本 PR 是该升级过程中为修复测试失败而做的配置调整