PR #42909 完整报告

vllm-project/vllm

[ROCm][CI] Stabilize ROCm pooling and multimodal CI

合并时间：2026-05-18 11:57

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42909

执行摘要

- 一句话：稳定 ROCm 池化与多模态 CI 测试
- 推荐动作：建议阅读以了解测试稳定性策略，特别是 assert_prompt_tokens 的设计和 ROCm 环境显式化方法。但 transformers/base.py 改动应等待进一步验证，合并后如有问题可关注 #42923。

功能与动机

来自 PR body："Addresses two ROCm CI groups that were failing or exposing cache-dependent behavior...The fixes keep model correctness checks intact. They avoid pinning model revisions where cache refreshes can legitimately change tokenizer/checkpoint details, and they make ROCm test setup explicit where defaults previously let unrelated code paths leak into the test."

实现拆解

1. 增强 token 计数断言的灵活性(test_max_tokens_per_doc.py)：引入 ExpectedPromptTokens 类型（int 或 tuple[int, ...]）和 assert_prompt_tokens 函数，允许在 expected 为元组时检查实际值是否在其中。为 intfloat/multilingual-e5-small 模型配置两种可能的 prompt token 数 (285/286 和 155/156)，避免因缓存不同导致 CI 失败。截断目标检查保持严格。
2. 显式化 ROCm 测试环境配置(test_gritlm.py)：在 test_gritlm_api_server_embedding 函数中从 tests.utils 导入 ROCM_EXTRA_ARGS 和 ROCM_ENV_OVERRIDES，注入 RemoteOpenAIServer 构造，确保 ROCm CI 使用正确的环境变量和额外参数，避免默认环境干扰。
3. 避免多模态测试中视频形状误用(test_qwen2_5_vl.py)：定义 IMAGE_ONLY_LIMIT_MM_PER_PROMPT = {"image": 1, "video": 0}，在图像窗口注意力测试中替代原来的 {"image": 1}，确保声明的能力与实际图像匹配，防止编码器预热时意外配置视频形状导致 GPU 挂起。
4. 添加 Gemma3 旧版 Checkpoint 的权重映射兼容性(transformers/base.py)：在 _create_hf_to_vllm_mapper 中检查模型是否具有 vision_tower 且无 vision_tower.vision_model 子属性，若成立则添加正则映射将 vision_tower.vision_model.* 重写为 model.vision_tower.*，使旧缓存 checkpoint 能正确加载。注意此改动存在争议（hmellor 认为可能错误并已提交 revert PR #42923）。

关键文件：

- tests/models/language/pooling/test_max_tokens_per_doc.py（模块 池化测试；类别 test；类型 test-coverage；符号 assert_prompt_tokens, ExpectedPromptTokens）：引入 ExpectedPromptTokens 类型和 assert_prompt_tokens 函数，使得测试能够容忍缓存导致的 token 计数微小差异，从而在不固定模型 revision 的情况下保持 CI 稳定。
- vllm/model_executor/models/transformers/base.py（模块 模型加载；类别 source；类型 data-contract；符号 _create_hf_to_vllm_mapper）：添加 Gemma3 旧版 checkpoint 的 vision_tower 路径映射，解决因 Transformers 版本差异导致的 key 不匹配问题。
- tests/models/language/pooling/test_gritlm.py（模块 GritLM 测试；类别 test；类型 test-coverage）：显式设置 ROCm 环境变量和额外参数，避免默认值泄露导致测试不稳定。
- tests/models/multimodal/generation/test_qwen2_5_vl.py（模块 多模态测试；类别 test；类型 test-coverage）：通过显式设置 video=0 确保图像测试不意外触发视频编码器形状，防止 GPU hang。

关键符号：assert_prompt_tokens, _create_hf_to_vllm_mapper

关键源码片段

tests/models/language/pooling/test_max_tokens_per_doc.py

引入 ExpectedPromptTokens 类型和 assert_prompt_tokens 函数，使得测试能够容忍缓存导致的 token 计数微小差异，从而在不固定模型 revision 的情况下保持 CI 稳定。

```
# 类型别名：可以是单个 int 或 int 元组
ExpectedPromptTokens = int | tuple[int, ...]

# 断言函数：如果 expected 是元组则检查 actual 在其中
def assert_prompt_tokens(actual: int, expected: ExpectedPromptTokens) -> None:
    if isinstance(expected, int):
        assert actual == expected
    else:
        assert actual in expected

# 在 bi-encoder 配置中使用元组，允许两种已知 prompt token 数
TestConfig(
    model="intfloat/multilingual-e5-small",
    args=["--enforce-eager", "--max-model-len", "512", "--trust-remote-code"],
    # 缓存差异导致两种可能值，保持截断检查严格
    without_truncated_prompt_tokens=(285, 286),
    with_max_tokens_per_query_prompt_tokens=(155, 156),
    with_max_tokens_per_doc_prompt_tokens=155,
    with_max_tokens_per_query_and_doc_prompt_tokens=25,
)
```

vllm/model_executor/models/transformers/base.py

添加 Gemma3 旧版 checkpoint 的 vision_tower 路径映射，解决因 Transformers 版本差异导致的 key 不匹配问题。

```
# 在 _create_hf_to_vllm_mapper 中，处理完常规重命名后：
# Gemma3 checkpoints saved with older Transformers versions include an
# extra vision_model level that the current AutoModel no longer has.
vision_tower = getattr(self.model, "vision_tower", None)
if vision_tower is not None and not hasattr(vision_tower, "vision_model"):
    # 将 vision_tower.vision_model.* 重映射到 model.vision_tower.*
    orig_to_new_regex[
        re.compile(r"^(?:model\.)?vision_tower\.vision_model\.(.+)")
    ] = r"model.vision_tower.\1"
```

评论区精华

讨论 1：正则表达式改进建议

- 参与者：gemini-code-assist[bot]
- 要点：建议将 regex 改为更通用形式以处理任意模型前缀（如 transformer. 或 gemma.），而非仅固定 model. 前缀。
- 结论：未采纳，PR 已合并但后续可能被 revert。

讨论 2：vision_tower 映射的质疑

- 参与者：hmellor
- 要点：认为 Gemma3 checkpoint 一年未变，该映射不必要且可能错误，已提交 revert PR #42923。
- 结论：已合并但计划 revert，争议未解决，需要跟踪。
- 正则表达式改进建议 (correctness): 未采纳，PR 已合并但后续可能被 revert。
- vision_tower 映射的质疑 (correctness): 已合并但计划 revert，争议未解决。

风险与影响

- 风险：主要风险来自 transformers/base.py 的改动：可能不正确地改变其他模型的权重加载路径，hmellor 已提出 revert，影响面不确定。测试参数放宽（允许多个 token 计数值）可能掩盖真实回归，但截断目标检查仍严格，风险较低。
- 影响：对 ROCm CI 稳定性有正面影响，但 Gemma3 模型在部分环境下可能出现加载失败（如果 revert 前），用户需关注。其他平台用户不受影响。测试改动均为 ROCm 专用，不影响默认流程。
- 风险标记：模型加载路径变更，兼容性争议，测试参数放宽

关联脉络

- PR #42923 Revert vision_tower mapping: hmellor 认为该改动不正确并已开此 PR 回滚。