

PR #52122 完整报告

vllm-project/vllm

[Bugfix][MiniCPM-V] Fix AssertionError in get_dummy_mm_data when passing VideoDummyOptions to _get_dummy_images

合并时间: 2026-08-13 23:44

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52122>

执行摘要

- 一句话: 修复 MiniCPM-V 视频 dummy 数据构建断言错误
- 推荐动作: 该 PR 是值得关注的 bugfix, 但代码改动较小, 逻辑直接, 不建议精读。关注点在于如何处理不同平台的测试跳过, 以及 dummy 数据生成中类型契约的维护。潜在改进方向包括: 为 _get_dummy_images 的断言提供更清晰的错误信息, 或者增加对不同平台测试的覆盖。

功能与动机

该 PR 解决了 `pytest tests/lora/test_minicpmv_tp.py::test_minicpmv_lora` 在非 CUDA 平台 (如 XPU) 上抛出的 `AssertionError`。根本原因是 commit `9a276d6375` 在 `dummy_inputs.py` 的 `_get_dummy_images` 中增加了运行时断言: `assert overrides is None or isinstance(overrides, ImageDummyOptions)`。然而 `MiniCPMVDummyInputsBuilder.get_dummy_mm_data` 在构造视频的逐帧 dummy 图像时, 一直直接传递 `video_overrides` (类型为 `VideoDummyOptions`)。此前该类型不匹配被静默忽略, 断言引入后直接触发异常。上游 CI 未能发现是因为 `test_minicpmv_lora` 使用了 `@pytest.mark.skipif(current_platform.is_cuda_alike(), ...)` 装饰器, 在 CUDA CI 上被跳过。

实现拆解

本次变更集中在一个文件 `vllm/model_executor/models/minicpmv.py` 的 `MiniCPMVDummyInputsBuilder.get_dummy_mm_data` 方法中。

1. 调整导入语句: 在 `vllm/config/multimodal` 的导入中增加 `ImageDummyOptions` 和 `VideoDummyOptions`, 确保类型判断和转换可用。
2. 在 `get_dummy_mm_data` 内部, 将原本直接传递给 `_get_dummy_images` 的 `video_overrides` 转换为新的 `video_frame_overrides` 变量。具体逻辑:
 - 若 `video_overrides` 是 `VideoDummyOptions` 实例, 则计算有效的 `num_video_frames`, 取原值和 `override` 中 `num_frames` 的较小值。
 - 若 `override` 中指定了 `width` 或 `height`, 则构造一个新的 `ImageDummyOptions` 实例, 保留这两个字段, 同时丢弃 `num_frames` (因为 `num_frames` 已单独处理)。
3. 调用 `_get_dummy_images` 时传入 `video_frame_overrides`, 而不是原来的 `video_overrides`, 确保参数类型匹配新的断言。

该修改保证了 `_get_dummy_images` 的契约一致性，同时保留了视频帧数量和宽高的覆盖能力。由于没有新增测试文件，对于 XPU 等平台的回归覆盖仍依赖于既有测试，但本次修复正是为了让该测试在这些平台上得以通过，因此可视为对既有测试的隐式修复。

关键文件：

- `vllm/model_executor/models/minicpmv.py`（模块 模型；类别 source；类型 data-contract；符号 `MiniCPMVDummyInputsBuilder.get_dummy_mm_data`）：该文件包含 `MiniCPMVDummyInputsBuilder` 的 `get_dummy_mm_data` 方法，是 `AssertionError` 的源头。修复在此文件中进行了类型转换和帧数覆盖逻辑，是本次变更的核心。

关键符号：`MiniCPMVDummyInputsBuilder.get_dummy_mm_data`

关键源码片段

`vllm/model_executor/models/minicpmv.py`

该文件包含 `MiniCPMVDummyInputsBuilder` 的 `get_dummy_mm_data` 方法，是 `AssertionError` 的源头。修复在此文件中进行了类型转换和帧数覆盖逻辑，是本次变更的核心。

```
# vllm/model_executor/models/minicpmv.py
# 该方法构建 dummy 多模态数据，包含图片和视频的占位输入。
class MiniCPMVDummyInputsBuilder(BaseDummyInputsBuilder[I]):
    def get_dummy_mm_data(
        self,
        seq_len: int,
        mm_counts: Mapping[str, int],
        mm_options: Mapping[str, BaseDummyOptions],
    ) -> MultiModalDataDict:
        num_images = mm_counts.get("image", 0)
        num_videos = mm_counts.get("video", 0)

        image_width, image_height = self.info.get_image_size_with_most_features()
        video_width, video_height = self.info.get_video_frame_size_with_most_features()
        num_video_frames = self.info.get_num_frames_with_most_features(
            seq_len, mm_counts
        )

        image_overrides = mm_options.get("image")
        video_overrides = mm_options.get("video")

        # 将视频的 overrides 转换为图像 overrides，以便逐帧生成 dummy 图像时
        # 与 _get_dummy_images 期望的 ImageDummyOptions 类型匹配。
        video_frame_overrides: ImageDummyOptions | None = None
        if isinstance(video_overrides, VideoDummyOptions):
            # 应用 num_frames 覆盖，取较小值以避免超出预算。
            if video_overrides.num_frames:
                num_video_frames = min(num_video_frames, video_overrides.num_frames)
            # 宽高覆盖时构造新的 ImageDummyOptions，仅保留尺寸字段。
            if video_overrides.width or video_overrides.height:
```

```

        video_frame_overrides = ImageDummyOptions(
            width=video_overrides.width,
            height=video_overrides.height,
        )

    return {
        "image": self._get_dummy_images(
            width=image_width,
            height=image_height,
            num_images=num_images,
            overrides=image_overrides,
        ),
        "video": [
            self._get_dummy_images(
                width=video_width,
                height=video_height,
                num_images=num_video_frames,
                # 这里原本直接传 video_overrides, 现在改为转换后的
                # video_frame_overrides, 以符合 _get_dummy_images 的断言。
                overrides=video_frame_overrides,
            )
        ] * num_videos,
    }

```

评论区精华

该 PR 的 review 讨论主要围绕自动化审核工具 (claude[bot]) 的评论展开。claude[bot] 指出这是一个 fork 的 PR，自动审核被禁用，需要维护者手动评论 @claude review 以触发一次性审查。随后维护者 jikunshang 和 DarkLight1337 分别批准了该 PR，未留下额外的技术评论。因此，讨论主要关注流程而非代码实现。

- 自动化 review 禁用 (process): 维护者未触发额外审查，但 jikunshang 和 DarkLight1337 已批准该 PR。

风险与影响

- 风险：该改动虽小，但涉及 dummy 数据生成的核心路径。风险点包括：
 1. 行为变化：通过应用 num_frames 覆盖来限制 num_video_frames，可能会改变视频帧数计算，从而影响 dummy 数据的 shape。虽然这是预期行为，但需要确认下游处理能正确处理较小的帧数。
 2. 类型转换：若 video_overrides 不是 VideoDummyOptions 而是其他类型（如通过 mm_options 键误传 image override），则 video_frame_overrides 保持 None，行为与之前一致，缺乏显式警告。
 3. 测试覆盖：本 PR 没有直接修改测试文件，对于非 CUDA 平台的验证依赖于既有测试，但这些测试在 CI 中被跳过，因此无法在 CI 中自动验证该修复。建议在本地或专门的 XPU CI 上运行相关测试。- 影响：影响范围较小，仅涉及 MiniCPM-V 模型在非 CUDA 平台（如 XPU、CPU）上生成 dummy 数据的功能。修复后，这些平台上的 LoRA 测试

不再因断言错误而失败，提升了测试兼容性。对 CUDA 平台用户无影响，因为该路径在 CUDA 上不会被触发（测试被跳过）。团队内受益于跨平台的测试稳定性。 - 风险标记：非 CUDA 平台测试覆盖不足，类型断言可能遗漏其他模型

关联脉络

- PR #9 commit 9a276d6375 (not in history): 该 PR 修复的断言正是在此 commit 中引入的，是本次 bugfix 的触发原因。
- PR #48608 [Bugfix] Video loading: sample over presentable frames, not header sample count (MP4 edit-list trims): 同属多模态视频处理修复，虽然针对加载逻辑，但可能涉及 dummy 数据生成的相关路径。