

PR #56385 完整报告

vllm-project/vllm

[Bugfix][MM] Fix swapped H/W in dummy video profiling inputs

合并时间: 2026-09-11 13:25

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/56385>

执行摘要

- 一句话: 修复多模态 dummy video 宽高转置
- 推荐动作: 不必逐行精读, 但值得花 5 分钟了解此处 "形状约定不一致" 的经典案例: 真正有价值的是认识到 profiling dummy 输入与真实数据处理路径之间存在隐式契约, 正方形默认值会长期掩盖 H/W 转置。阅读时重点关注三点: (1) 为什么以 vllm/multimodal/video.py 的 `rescale_video_size` 作为约定基准; (2) dummy 形状变化会经由 token 估算传导到 KV cache 与 `max_num_seqs`; (3) 该修复缺少单测, 若后续要动 `_get_dummy_videos` 或各模型的 `VideoDummyOptions`, 建议先补一条 `shape == (T, H, W, C)` 的断言测试。

功能与动机

PR body 明确指出动机: `BaseDummyInputsBuilder._get_dummy_videos` 构造 dummy profiling video 时把宽度和高度写反了, "for any non-square dummy size, the profiling video has H and W transposed, feeding memory profiling / max-token estimation the wrong aspect ratio", 且 "Square sizes hide the bug"。仓库内其它视频路径 (如 `vllm/multimodal/video.py` 的 `rescale_video_size` 用 `_, height, width, _ = frames.shape`) 都按 (T, H, W, C) 处理, 兄弟方法 `_get_dummy_images` 也已使用正确顺序 (`Image.new("RGB", (width, height))`), 因此这是一处与全局约定背离的隐性错误, 需要在进入真实 profiling 计算前修正。

实现拆解

1. 定位入口: 变更集中在 `vllm/multimodal/processing/dummy_inputs.py` 的 `BaseDummyInputsBuilder._get_dummy_videos`, 该方法是多模态模型启动 profiling 时生成 dummy 视频样本的唯一入口, 约在函数末尾用 `np.full(...)` 一次性构造整段视频张量。
2. 对齐布局约定: 全仓库视频数组统一为 (T, H, W, C), 证据是 `vllm/multimodal/video.py` 的 `rescale_video_size` 按 `_, height, width, _ = frames.shape` 解析高与宽; `_get_dummy_images` 采用 `Image.new("RGB", (width, height))` 的 PIL 顺序 (W, H) 也等价于 height 在前的语义。因此修复只改数组形状元组中的 H/W 位置, 不改动 `overrides` 对 `num_frames / width / height` 的 min 收缩与 warning 分支。
3. 影响链路: dummy video 会经模型 processor 的 `resize` 与 `patchify` 流程折算为视觉 token 数, 参与 memory profiling 的峰值激活估算, 进而决定 KV cache 预算与 `max_num_seqs` 等启动期参数。宽高转置会改变 smart-resize 后的实际形状, 使估算结果偏离真实请求的形状分布; 修复后估算与实际推理路径的输入布局一致。

4. 测试与配套：本 PR 只改 1 个文件、+1/-1，未新增任何单元测试或测试数据；作者在 PR body 中提供了使用真实代码路径的最小复现脚本（断言 `v.shape == (4, 1080, 1920, 3)`，main 上失败、修复后通过），并执行了 `pre-commit run ruff-check --files vllm/multimodal/processing/dummy_inputs.py`。
5. CI 与合入流程：maintainer DarkLight1337 评论 `/ci run` 触发 Buildkite CI #88244，随后直接 approve（回复 "Thanks"），无额外 review 评论；PR body 注明 "AI assistance was used to locate, reproduce, and fix this bug; a human has reviewed the change"，commit 带 Co-authored-by: Cursor Agent 标记。

关键文件：

- `vllm/multimodal/processing/dummy_inputs.py`（模块 多模态输入；类别 source；类型 core-logic；符号 `BaseDummyInputsBuilder._get_dummy_videos`）：本 PR 唯一改动文件。`BaseDummyInputsBuilder._get_dummy_videos` 生成的 dummy video 布局由 (T, W, H, C) 修正为 (T, H, W, C)，直接决定多模态模型 profiling 时的视频 token 估算与显存预算是否与真实输入一致。

关键符号：`BaseDummyInputsBuilder._get_dummy_videos`

关键源码片段

`vllm/multimodal/processing/dummy_inputs.py`

本 PR 唯一改动文件。`BaseDummyInputsBuilder._get_dummy_videos` 生成的 dummy video 布局由 (T, W, H, C) 修正为 (T, H, W, C)，直接决定多模态模型 profiling 时的视频 token 估算与显存预算是否与真实输入一致。

```
# vllm/multimodal/processing/dummy_inputs.py
# BaseDummyInputsBuilder._get_dummy_videos: 为 profiling 生成占位视频张量。
# 这里的形状语义必须与 vllm/multimodal/video.py 中 rescale_video_size 的
# `__, height, width, _ = frames.shape` 完全一致，否则下游会拿到转置后的宽高比。
def _get_dummy_videos(
    self,
    width: int,
    height: int,
    num_frames: int,
    num_videos: int,
    overrides: VideoDummyOptions | None = None,
) -> list[npt.NDArray]:
    if num_videos == 0:
        return []

    # overrides 只做「向下收缩」：显式值超过模型上限时打 warning 并忽略，
    # 因此越界场景下 width / height / num_frames 仍保持模型侧默认值。
    # 这段逻辑本次未改动，但它决定了最终进入 np.full 的 height / width。
    if overrides:
        if overrides.num_frames:
            if overrides.num_frames > num_frames:
                logger.warning(
```

```

        "video.num_frames override (%d) exceeds model's "
        "maximum number of frames (%d), will be ignored",
        overrides.num_frames,
        num_frames,
    )
    num_frames = min(num_frames, overrides.num_frames)
if overrides.width:
    if overrides.width > width:
        logger.warning(
            "video.width override (%d) exceeds model's "
            "maximum width (%d), will be ignored",
            overrides.width,
            width,
        )
    width = min(width, overrides.width)
if overrides.height:
    if overrides.height > height:
        logger.warning(
            "video.height override (%d) exceeds model's "
            "maximum height (%d), will be ignored",
            overrides.height,
            height,
        )
    height = min(height, overrides.height)

```

本次唯一改动：np.full 的形状从 (T, W, H, C) 变成 (T, H, W, C)。
 # 旧写法在非正方形 dummy 尺寸下会把高与宽转置（例如 1920x1080 被建成
 # (4, 1920, 1080, 3)），使 profiling 与 max-token 估算拿到错误宽高比；
 # 正方形尺寸会掩盖该问题。修改后与 rescale_video_size 的解析顺序一致。
 video = np.full((num_frames, height, width, 3), 255, dtype=np.uint8)

注意：num_videos 份 dummy 共享同一个 NDArray 对象，下游若有原地写需谨慎。
 return [video] * num_videos

评论区精华

本次讨论非常轻量，几乎没有技术交锋：claude[bot] 仅说明 fork PR 的自动 review 被禁用，需要 maintainer 手动 @claude review 才会跑一次性评审，实际未触发；maintainer DarkLight1337 只做了 /ci run 与一句 "Thanks" 的 approve，没有对形状约定或影响面提出质疑。作者则在 PR body 中承担了主要论证工作：用 rescale_video_size 的 _, height, width, _ = frames.shape 和 _get_dummy_images 的 Image.new("RGB", (width, height)) 两条证据说明 (T, H, W, C) 是既定约定，并附上可运行的复现脚本与前后对照输出（shape 由 (4, 1920, 1080, 3) 变为 (4, 1080, 1920, 3)）。整体结论是无争议的一行修复，但评审过程未产生关于 profiling token 数变化幅度的实测数据，也没有要求补测试。

- fork PR 的自动化 review 被禁用 (other): 自动化静态 / 语义评审缺席，正确性判断主要依赖作者在 PR body 提供的复现脚本以及 maintainer 的人工确认。

- (T, H, W, C) 布局约定的一致性确认 (correctness): 确认 (T, H, W, C) 为仓库内视频数组统一约定, 修复方向与范围均被接受, 无争议点。
- CI 验证范围与测试缺口 (testing): 合入仅依赖端到端 CI 与人工判断, 仓库内没有直接断言 dummy video 形状 of 的测试用例, 回归防护存在空缺。

风险与影响

- 风险:
 1. profiling 结果变化面: 只有 dummy video 尺寸非正方形 (width $\neq$ height, 如 16:9 配置) 的多模态模型受影响, 正方形配置行为完全不变; 修复后 token 数与激活峰值估算可能较旧值下降, 原先 " 偏保守 " 的显存预算会变得紧凑, 若下游还存在其它低估因素, 理论上增加 OOM 或 max_num_seqs 需回退的风险。
 2. 缺少测试覆盖: 改动落在 vllm/multimodal/processing/dummy_inputs.py, 但 PR 未在 tests/multimodal/ 下补任何形状断言, 回归只能依赖 CI 端到端视频用例, 后续重构很容易再次写反 H/W 而不被发现。
 3. 依赖隐式约定: 正确性依赖 vllm/multimodal/video.py 等下游模块持续按 (T, H, W, C) 解释张量, 而 dummy 张量本身不带元数据, 约定一旦漂移就会静默错算, 这类 " 形状语义靠注释与惯例维持 " 的结构风险未被本 PR 处理。
 4. 既有行为未变更点: return [video] * num_videos 让多个 dummy 视频共享同一个 NDArray 对象, 若下游有原地写操作会互相污染; 本次未改动该逻辑, 属于既有隐患, 仅作参考。
 5. 无真实推理回归风险: dummy 输入只用于 profiling, 不进入线上推理路径, 因此不会改变实际请求的输出质量或吞吐, 风险集中在启动期的显存与并发参数估算。- 影响: 用户侧: 使用视频输入的多模态模型 (视频理解类 VLM) 在非正方形 dummy video 配置下, 启动 profiling 得到的显存 / 并发参数会更贴近真实输入形状, 降低因估算偏差导致的启动即 OOM 或并发被过度压低的概率。系统侧: 影响 memory profiling $\rightarrow$ KV cache 预算 $\rightarrow$ max_num_seqs 这条启动关键链路, 但改动是单点、单文件、+1/-1, 不会扩散到运行时调度或注意力后端。团队侧: 给多模态维护者一个明确信号——dummy 输入布局必须与模型 processor 的 (T, H, W, C) 约定保持一致, 后续新增视频模型 (自定义 VideoDummyOptions) 时应显式检查 width/height 顺序。整体影响范围窄、程度中等, 属于低扩散但落在启动路径上的正确性修复。- 风险标记: profiling 路径变更, 缺少单元测试覆盖, 隐式形状约定依赖, 非方形 dummy 尺寸行为变化

关联脉络

- PR #56310 [Bugfix][Multimodal] Restore cached audio inputs with UUIDs: 同属 vllm/multimodal/ 输入处理路径的 bugfix, 均修正多模态输入在解析 / 构造阶段因约定不一致导致的静默错误, 可以与本 PR 一起观察多模态输入层的稳定性趋势。
- PR #56070 [Bugfix][Frontend] Check EC requirements for each metadata item: 同一批多模态相关缺陷修复, 涉及多模态元数据校验路径; 与本 PR 共同反映近期多模态输入校验 / 构造逻辑的补漏节奏。
- PR #56335 [CI/Build] Pin HyperCLOVAX V2 test model revision: 多模态测试基础设施改动 (tests/models/registry.py)。本 PR 未新增测试, 若后续补 _get_dummy_videos 形状

单测，大概率会与该类多模态测试资产协同演进。