

PR #49190 完整报告

vllm-project/vllm

[bugfix] Fix Cosmos3 Edge checkpoint weights filtering, video loading, prompt expansion

合并时间: 2026-07-21 17:18

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49190>

执行摘要

- 一句话: 修复 Cosmos3 Edge 权重过滤、视频加载和占位符扩展
- 推荐动作: 值得阅读, 展示了多模态模型集成中常见的三类修复模式 (权重映射、后端复用、预处理对齐), 对类似模型接入有参考价值。测试覆盖充分, 可视为 bugfix 的典范。

功能与动机

Cosmos3 Edge 的 checkpoint 中包含生成器专属的 `k_norm_und_for_gen` 权重, 需要跳过以免干扰主干模型加载; 视频处理器需要与 Qwen3-VL 共享视频后端以避免重复实现; 占位符扩展需要产生正确的逐帧 wrapper 以便下游正确处理。

实现拆解

1. 权重过滤: 在 `vllm/model_executor/models/cosmos3_edge.py` 的 `orig_to_new_substr` 字典中添加 `"k_norm_und_for_gen": None`, 使权重映射时忽略该键。
2. 视频后端路由: 在 `vllm/multimodal/video.py` 中将 `Qwen3VLVideoBackend` 的 `video_processor` 参数从单字符串改为元组 (`"Qwen3VLVideoProcessor", "Cosmos3EdgeVideoProcessor"`), 使 Cosmos3 Edge 的视频预处理复用同一后端。
3. 占位符扩展修正: 在 `Cosmos3EdgeMultiModalProcessor` 中新增静态方法 `_expands_only_video_token` 返回 `False`, 指示处理器将整个视觉占位符替换为带时间戳的逐帧序列, 而不仅是替换 `<lvideo_padl>` 标记。
4. 测试覆盖: 更新 `tests/models/multimodal/processing/test_cosmos3_edge.py` 中 `_assert_video_outputs` 验证帧 wrapper 数量; 新增 `tests/multimodal/test_video.py` 中的 `test_cosmos3_edge_uses_qwen3_vl_video_backend` 确认后端路由; 在 `tests/models/multimodal/test_mapping.py` 中添加被跳过的权重。

关键文件:

- `vllm/model_executor/models/cosmos3_edge.py` (模块 模型定义; 类别 `source`; 类型 `data-contract`; 符号 `_expands_only_video_token`): 核心修复文件: 添加权重过滤规则和占位符扩展控制方法。
- `tests/models/multimodal/processing/test_cosmos3_edge.py` (模块 模型测试; 类别 `test`; 类型 `test-coverage`): 增强视频输出断言, 验证帧 wrapper 数量。
- `tests/multimodal/test_video.py` (模块 视频测试; 类别 `test`; 类型 `test-coverage`; 符号 `test_cosmos3_edge_uses_qwen3_vl_video_backend`): 新增测试确保

Cosmos3EdgeVideoProcessor 被正确路由到 Qwen3VLVideoBackend。

- vllm/multimodal/video.py (模块 视频处理; 类别 source; 类型 core-logic) : 注册路由: 将 Cosmos3EdgeVideoProcessor 加入 Qwen3VLVideoBackend 的处理范围。
- tests/models/multimodal/test_mapping.py (模块 映射测试; 类别 test; 类型 test-coverage) : 更新权重映射测试: 增加 k_norm_und_for_gen.weight 作为应被跳过的权重条目。

关键符号: Cosmos3EdgeMultiModalProcessor._expands_only_video_token, test_cosmos3_edge_uses_qwen3_vl_video_backend

关键源码片段

vllm/model_executor/models/cosmos3_edge.py

核心修复文件: 添加权重过滤规则和占位符扩展控制方法。

```
class Cosmos3EdgeMultiModalProcessor(Qwen3VLMultiModalProcessor):
    @staticmethod
    def _expands_only_video_token(_hf_processor: ProcessorMixin) -> bool:
        # Cosmos 将每个视频渲染为一个完整的
        # <lvision_startl><lvideo_padi><lvision_endl> 占位符,
        # 其参照处理器会将整个三元组替换为带时间戳的逐帧视觉序列。
        # 返回 False 表示不 ** 只 ** 扩展 video_pad, 而是扩展整个三元组。
        return False

...
hf_to_vllm_mapper = WeightsMapper(
    orig_to_new_substr={
        "_moe_gen": None,
        "k_norm_und_for_gen": None, # 跳过生成器专属权重, 避免干扰主干加载
        ...
    },
    ...
)
```

评论区精华

无实质性 reviewer 讨论。仅有一条 Claude code review 自动评论 (因来自 fork 未自动审查), 最终由维护者 Isotr0py 批准合并。

- 暂无高价值评论线程

风险与影响

- 风险: 风险较低。权重过滤仅忽略特定 key, 不影响其他权重; 视频后端路由为现有模块扩展, 不修改其他注册项; 占位符扩展行为通过新增静态方法控制, 不改变父类默认逻辑。潜在风险是若未来 checkpoint 中包含类似的其他生成器专属权重而未被过滤, 可能导致加载失败, 但已通过测试覆盖了当前已知的权重。

- 影响：影响范围限于 Cosmos3 Edge 模型用户。修复后：checkpoint 加载正确不再因未知权重报错；视频解码复用 Qwen3-VL 的后端，确保一致性和性能；占位符扩展产生正确的逐帧标记序列，使视频输入处理符合模型预期。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR