

PR #48912 完整报告

vllm-project/vllm

[Model] Enable EVS for Qwen3.5

合并时间: 2026-07-27 21:42

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48912>

执行摘要

- 一句话: 为 Qwen3.5 启用 EVS 视频采样优化
- 推荐动作: 值得合并, 因为变更范围小, 有明确的基准测试验证, 且经过充分讨论。建议在未来 PR 中为 Qwen3.5 的 EVS 添加专门的集成测试, 确保父类 EVS 逻辑变更时不会破坏功能。

功能与动机

此前 Qwen3.5 对 EVS 硬编码禁用 (`supports_multimodal_pruning = False`), 而 EVS 是 Qwen3-VL 父类支持的有效优化。PR body 中的基准测试表明 EVS 在 Qwen3.5 上有效, 准确率下降 (0.9%) 和 TTFT 提升 (31.8%) 与 Qwen3-VL 相当, 因此有充分的理由启用。

实现拆解

本 PR 仅修改 `vllm/model_executor/models/qwen3_5.py` 一个文件, 关键变更包括:

1. 启用 EVS 标志: 将类属性 `supports_multimodal_pruning` 从 `False` 改为 `True`。
2. 动态配置 EVS 启用: 在 `__init__` 中将 `is_multimodal_pruning_enabled` 从固定 `False` 改为通过 `multimodal_config.is_multimodal_pruning_enabled()` 动态获取。
3. 添加 EVS 所需属性: `video_pruning_rate` (来自配置) 和 `_tokenizer` (通过 `cached_tokenizer_from_config`)。
4. 添加父类所需 `deepstack` 属性: 由于 EVS 函数继承自 `Qwen3VLForConditionalGeneration`, 需要设置 `use_deepstack`、`deepstack_num_level`、`visual_dim` 和 `multiscale_dim` (Qwen3.5 不使用 `deepstack`, 这些属性设为默认值 / 零)。
5. 删除禁止 EVS 的方法: 移除了引发 `NotImplementedError` 的 `recompute_mrope_positions` 方法, 该守卫已不再需要。
6. 更新导入: 添加 `from vllm.tokenizers.registry import cached_tokenizer_from_config`。

以上变更在 Qwen3.5 的 `dense` 和 `MoE` 两个模型类中对称进行 (`Qwen3_5ForConditionalGeneration` 和 `Qwen3_5MoeForConditionalGeneration` 的 `__init__` 方法)。

关键文件:

- `vllm/model_executor/models/qwen3_5.py` (模块 模型执行器; 类别 `source`; 类型 `data-contract`; 符号 `recompute_mrope_positions`): 唯一修改的文件; 启用了 EVS 并添

加了必要的属性初始化

关键符号：Qwen3_5ForConditionalGeneration.init,
Qwen3_5MoeForConditionalGeneration.init

关键源码片段

vllm/model_executor/models/qwen3_5.py

唯一修改的文件；启用了 EVS 并添加了必要的属性初始化

```
# Qwen3_5ForConditionalGeneration 启用 EVS 的关键变更
# 文件：vllm/model_executor/models/qwen3_5.py

# 新增导入
from vllm.tokenizers.registry import cached_tokenizer_from_config

# 类属性变更：支持多模态剪枝（EVS）
class Qwen3_5ForConditionalGeneration(Qwen3VLForConditionalGeneration, IsHybrid):
    supports_multimodal_pruning = True # 之前为 False

    def __init__(self, *, vllm_config: VllmConfig, prefix: str = "model"):
        # ... 原有初始化 ...
        # 动态启用 EVS（而非硬编码 False）
        self.is_multimodal_pruning_enabled = (
            multimodal_config.is_multimodal_pruning_enabled()
        )
        self.video_pruning_rate = self.multimodal_config.video_pruning_rate
        self._tokenizer = cached_tokenizer_from_config(vllm_config.model_config)

        # 以下属性是 EVS 函数从父类继承所必需的
        # Qwen3.5 不使用 deepstack，因此设置为默认值
        self.use_deepstack = hasattr(config.vision_config, "deepstack_visual_indexes")
        self.deepstack_num_level = (
            len(config.vision_config.deepstack_visual_indexes)
            if self.use_deepstack
            else 0
        )
        self.visual_dim = config.vision_config.out_hidden_size
        self.multiscale_dim = self.visual_dim * self.deepstack_num_level

        # 移除了以下方法，因为 EVS 现已支持
        # def recompute_mrope_positions(self, *args, **kwargs):
        #     raise NotImplementedError("Qwen3.5 does not support multimodal pruning (EVS).")
```

评论区精华

Reviewer Isotr0py 最初评论说 Qwen3.5 不使用 deepstack，暗示移除 deepstack 相关属性。作者 garrygale 解释说这些属性是 EVS 函数从父类继承所必需的，移除会导致错误。Isotr0py 随后要求添加注释说明这些属性仅用于 EVS，作者照做了，最终获得批准。这反映了

代码复用与设计明确性之间的权衡。

- deepstack 属性必要性 (design): 保留 deepstack 属性, 并添加注释说明仅用于 EVS。
- 代码注释要求 (documentation): 已添加注释说明仅用于 EVS。

风险与影响

- 风险:
 1. 回归风险 (低): 只涉及模型文件本身, EVS 逻辑由父类提供, Qwen3.5 模型自身的推理和训练不受影响。
 2. 兼容性风险 (低): 添加的属性都有默认值 (如 deepstack_num_level=0), 不影响已有调用路径。
 3. 性能风险 (无): EVS 本身是性能优化, 仅在启用时生效。
 4. 测试覆盖风险 (中): 没有新增测试用例, 依赖父类的 EVS 测试。如果父类 EVS 逻辑变更, Qwen3.5 可能被遗漏。
- 影响:
 - 用户: 使用 Qwen3.5 的用户现在可以启用 EVS (通过 --video-pruning-rate) 获得 TTFT 优化, 适合视频推理场景。
 - 系统: 无系统级影响; 模型配置和行为完全向后兼容。
 - 团队: 变更简洁 (32 行增加, 12 行删除), 易于维护。
 - 影响程度: 低
 - 风险标记: 缺少测试覆盖

关联脉络

- 暂无明显关联 PR