PR #46108 完整报告

vllm-project/vllm

[Model] ColQwen3.5: fix retrieval correctness (bias + bidirectional)

合并时间：2026-06-22 17:25

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/46108

执行摘要

- 一句话：修复 ColQwen3.5 检索投影偏置与双向注意力问题
- 推荐动作：值得精读，尤其是 config.py 的数据契约设计和注意力类型处理。对多模态检索模型维护者有参考价值。

功能与动机

The in-tree ColQwen3_5 model deviates from the native colpali ColQwen3_5Processor inference pipeline in three silent ways, none caught by the current sanity-level tests. Measured cost: ~2.5 ndcg@10 on Vidore3.

实现拆解

步骤拆解：

1. 修复投影偏置(vllm/model_executor/models/colqwen3_5.py)：将 custom_text_proj 从 bias=False 改为 bias=True 并零初始化，使训练好的偏置能被 load_weights 加载。既兼容无偏置 checkpoint，也正确加载有偏置的 checkpoint。
2. 启用双向注意力(config.py, qwen3_next.py, attention.py)：新建 ColQwen3_5Config 继承 Qwen3.5 配置，设置 hf_config.is_causal=False；Qwen3NextAttention 根据 is_causal 选择 AttentionType.ENCODER_ONLY 并传入 attn 参数；Attention.get_kv_cache_spec 对 ENCODER_ONLY 返回 None，避免混合模型中 runner 遍历时断言失败。
3. 配套文档与测试：新增配置验证测试 test_colqwen3_5_config_enables_bidirectional_attention，更新示例脚本说明视觉 token 预算和 prompt 模板，更新模型文档。

关键文件：

- vllm/model_executor/models/config.py（模块 模型配置；类别 source；类型 data-contract；符号 ColQwen3_5Config, verify_and_update_model_config）：引入了 ColQwen3_5Config，设置 is_causal=False，并更新 MODELS_CONFIG_MAP 映射，是双向注意力的配置基础。
- vllm/model_executor/models/qwen3_next.py（模块 注意力层；类别 source；类型 data-contract；符号 Qwen3NextAttention, AttentionType）：修改 Qwen3NextAttention，根据 config.is_causal 选择 AttentionType.ENCODER_ONLY 而非默认 DECODER，实现双向注意力。

- vllm/model_executor/models/colqwen3_5.py（模块 模型定义；类别 source；类型 data-contract）：修复 custom_text_proj 偏置，从 bias=False 改为 bias=True 并零初始化，确保训练偏置可加载。
- vllm/model_executor/layers/attention/attention.py（模块 注意力后端；类别 source；类型 data-contract；符号 get_kv_cache_spec）：修改 get_kv_cache_spec 对 ENCODER/ENCODER_ONLY 类型返回 None，避免混合模型中 runner 断言失败。
- tests/models/multimodal/pooling/test_colqwen3_5.py（模块 测试；类别 test；类型 test-coverage；符号 test_colqwen3_5_config_enables_bidirectional_attention）：新增测试验证 ColQwen3_5Config 正确设置 is_causal=False，无需 GPU。
- examples/pooling/score/colqwen3_5_rerank_online.py（模块 示例；类别 source；类型 core-logic）：更新示例，提供正确的处理器参数和 prompt 模板说明，确保用户能与原生管线匹配。
- docs/models/pooling_models/token_embed.md（模块 文档；类别 docs；类型 documentation）：添加 VultronRetrieverPrime-8B 到文档中。

关键符号：ColQwen3_5Config.verify_and_update_model_config, Qwen3NextAttention.init, Attention.get_kv_cache_spec, test_colqwen3_5_config_enables_bidirectional_attention

关键源码片段

vllm/model_executor/models/config.py

引入了 ColQwen3_5Config，设置 is_causal=False，并更新 MODELS_CONFIG_MAP 映射，是双向注意力的配置基础。

```
# ColQwen3.5 使用双向注意力，继承 Qwen3.5 的 mamba 缓存处理
class ColQwen3_5Config(Qwen3_5ForConditionalGenerationConfig):
    """ColQwen3.5 (late-interaction retrieval) inherits Qwen3.5's mamba cache
    handling and additionally serves BIDIRECTIONAL attention: ColPali-style
    document/query encoding attends over the whole sequence, not causally. Set
    is_causal=False so Qwen3NextAttention builds its full_attention layers with
    AttentionType.ENCODER_ONLY (the linear_attention GatedDeltaNet layers are
    unaffected). Generation arches keep the parent (causal) and are untouched.
    """

    @staticmethod
    def verify_and_update_model_config(model_config: "ModelConfig") -> None:
        model_config.hf_config.is_causal = False

# 在注册表中替换为新的配置类
MODELS_CONFIG_MAP["ColQwen3_5"] = ColQwen3_5Config # 原为 Qwen3_
5ForConditionalGenerationConfig
```

vllm/model_executor/models/qwen3_next.py

修改 Qwen3NextAttention，根据 config.is_causal 选择 AttentionType.ENCODER_ONLY 而非默认 DECODER，实现双向注意力。

```
# 在 Qwen3NextAttention.__init__ 中，根据 config.is_causal 选择注意力类型
# 默认为因果注意力（DECODER），若 is_causal=False 则使用双向编码器注意力（ENCODER_
ONLY）
attn_type = (
    AttentionType.DECODER
    if getattr(config, "is_causal", True)
    else AttentionType.ENCODER_ONLY
)
self.attn = Attention(
    self.num_heads,
    self.head_dim,
    self.scaling,
    num_kv_heads=self.num_kv_heads,
    cache_config=cache_config,
    quant_config=quant_config,
    prefix=f"{prefix}.attn",
    attn_type=attn_type, # 显式传入注意力类型，之前未传递，默认 DECODER
    # 其他参数（layer_idx, dual_chunk_attention_config 等）
)
```

评论区精华

无实质性审查讨论，仅 mergify 自动化操作和作者请求 ready 标签。

- 暂无高价值评论线程

风险与影响

- 风险：
  - 回归风险：生成模型不受影响（is_causal 未设置时默认 True，维持 DECODER 行为），仅 ColQwen3.5 触发新路径。
  - 兼容风险：偏置零初始化确保无偏置 checkpoint 工作正常；有偏置 checkpoint 正确加载。
  - 性能风险：ENCODER_ONLY 注意力不维护 KV 缓存，可能节省内存，但需实际评估。
  - 测试风险：新增配置测试不包含 GPU 推理覆盖，不能保证所有后端兼容。
- 影响：
  - 用户影响：ColQwen3.5 用户检索正确性显著提升，无需修改代码。
  - 系统影响：无。
  - 团队影响：明确了后续 ColPali 风格模型实现的注意事项。
  - 风险标记：核心路径变更（注意力类型），缺少 GPU 推理测试覆盖，兼容性（零初始化偏置）

关联脉络

- PR #36887 Add ColQwen3.5 late interaction model: 此 PR 是 #36887 的后续修复，修正其 ColQwen3.5 实现中的三个正确性缺陷。