PR #45131 完整报告

vllm-project/vllm

Deprecated 1st generation Qwen and QwenVL models

合并时间：2026-06-10 22:55

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45131

执行摘要

- 一句话：移除第一代 Qwen 和 QwenVL 模型代码及配套组件
- 推荐动作：此 PR 清理彻底，值得作为参考模板用于其他旧模型淘汰。建议阅读以理解 vLLM 的模型注册和组件解耦方式。如果用户仍依赖旧模型，不应升级到此版本。

功能与动机

PR 作者在描述中说明：'These models:

- have been superseded by several newer generations of Qwen architectures - are not used as building blocks for other architectures - incur maintenance burden. They may still be used in versions of vLLM <= 0.23.'

实现拆解

1. 删除核心模型文件：移除 vllm/model_executor/models/qwen.py 和 qwen_vl.py，这两个文件实现了 QWenModel、QWenAttention、VisualAttention 等核心类。
2. 删除自定义 tokenizer 和 processor：移除 vllm/tokenizers/qwen_vl.py 和 vllm/transformers_utils/processors/qwen_vl.py，删除了 QwenVLTokenizer、QwenVLProcessor 等。
3. 更新模型注册表：在 vllm/model_executor/models/registry.py 中移除 "QwenForCausalLM" 和 "QwenVLForConditionalGeneration" 条目。
4. 更新配置和模板注册：修改 vllm/config/model.py 移除相关配置项，调整 vllm/transformers_utils/chat_templates/registry.py 中的 fallback 逻辑。
5. 清理示例和测试：从 examples/generate/multimodal/ 下移除 Qwen-VL 的示例函数，并删除或修改相关测试 conftest 文件（如 tests/renderers/conftest.py、tests/tokenizers_/conftest.py、tests/models/multimodal/conftest.py），移除对旧模型的预缓存和引用。

关键文件：

- vllm/model_executor/models/qwen.py（模块 模型定义；类别 source；类型 deletion；符号 QWenMLP, QWenAttention, QWenBlock, QWenModel）：核心模型文件，包含 QWenModel、QWenAttention 等类，是整个 QWen 模型支持的基础，删除后彻底移除该架构。

- vllm/model_executor/models/qwen_vl.py（模块 多模态模型；类别 source；类型 deletion；符号 QwenImagePixelInputs, QwenImageEmbeddingInputs, VisualAttention, QwenVLMLP）：多模态模型文件，包含视觉注意力层、图像输入类型定义等，是 Qwen-VL 模型的核心，删除后移除整个多模态变体。
- vllm/tokenizers/qwen_vl.py（模块 分词器；类别 source；类型 deletion；符号 get_qwen_vl_tokenizer, TokenizerWithoutImagePad, tokenize, _decode）：自定义分词器文件，包含 QwenVLTokenizer 及其辅助函数，用于处理图像 token 的特殊逻辑，随模型一同移除。
- vllm/transformers_utils/processors/qwen_vl.py（模块 处理器；类别 source；类型 deletion；符号 QwenVLImageProcessorFast, QwenVLProcessor, init）：自定义 processor 文件，包含 QwenVLImageProcessorFast 和 QwenVLProcessor，用于图像预处理和 tokenizer/padding 集成，随模型移除。

关键符号：QWenMLP, QWenAttention, QWenBlock, QWenModel, QWenBaseModel, QwenImagePixelInputs, QwenImageEmbeddingInputs, VisualAttention, QwenVLMLP, VisualAttentionBlock, get_qwen_vl_tokenizer, TokenizerWithoutImagePad, QwenVLTokenizer, QwenVLImageProcessorFast, QwenVLProcessor

关键源码片段

vllm/model_executor/models/qwen.py

核心模型文件，包含 QWenModel、QWenAttention 等类，是整个 QWen 模型支持的基础，删除后彻底移除该架构。

```
# vllm/model_executor/models/qwen.py ( 已删除 )
# QWenMLP 是第一代 Qwen 模型的 MLP 层，使用 silu 激活和 MergedColumnParallelLinear。
# 由于该模型已被 Qwen2 及更新版本取代，整个文件被移除。
class QWenMLP(nn.Module):
    def __init__(
        self,
        hidden_size: int,
        intermediate_size: int,
        hidden_act: str = "silu",
        quant_config: QuantizationConfig | None = None,
        prefix: str = "",
    ):
        super().__init__()
        self.gate_up_proj = MergedColumnParallelLinear(
            hidden_size, [intermediate_size] * 2,
            bias=False, quant_config=quant_config,
            prefix=f"{prefix}.gate_up_proj",
        )
        self.c_proj = RowParallelLinear(
            intermediate_size, hidden_size,
            bias=False, quant_config=quant_config,
            prefix=f"{prefix}.c_proj",
        )
```

```
        if hidden_act != "silu":
            raise ValueError(
                f"Unsupported activation: {hidden_act}. Only silu is supported."
            )
        self.act_fn = SiluAndMul()

    def forward(self, x: torch.Tensor) -> torch.Tensor:
        gate_up, _ = self.gate_up_proj(x)
        x = self.act_fn(gate_up)
        x, _ = self.c_proj(x)
        return x
```

vllm/model_executor/models/qwen_vl.py

多模态模型文件，包含视觉注意力层、图像输入类型定义等，是 Qwen-VL 模型的核心，删除后移除整个多模态变体。

```
# vllm/model_executor/models/qwen_vl.py（已删除）
# 定义模型输入类型，使用 TensorSchema 做形状标注。
# 这些类型被用于多模态输入构建，现随整个模型被移除。
class QwenImagePixelInputs(TensorSchema):
    """像素值输入，形状 (bn, 3, h, w)"""
    type: Literal["pixel_values"] = "pixel_values"
    data: Annotated[torch.Tensor, TensorShape("bn", 3, "h", "w")]

class QwenImageEmbeddingInputs(TensorSchema):
    """图像嵌入输入，形状 (bn, 256, hs)，hs 需匹配语言模型隐藏层维度"""
    type: Literal["image_embeds"] = "image_embeds"
    data: Annotated[torch.Tensor, TensorShape("bn", 256, "hs")]

# 联合类型
QwenImageInputs: TypeAlias = QwenImagePixelInputs | QwenImageEmbeddingInputs
```

评论区精华

删除 QwenVLProcessor：Reviewer DarkLight1337 指出需要一并移除 QwenVLProcessor，作者在第二个 commit 中执行了移除，解决了该问题。 等待 PR#45127 合并：为了避免合并冲突，作者等待了相关的 ERNIE 模型移除 PR 合并后再操作。 pre-commit 检查：自动化检查提示 pre-commit 失败，作者后续修复。

- 需要移除 QwenVLProcessor (design): 作者在第二个 commit 中删除了 processor 文件，问题已关闭。

风险与影响

- 风险：向后兼容性：用户如果从旧版本升级到包含此 PR 的版本，使用 Qwen/QwenVL 模型的配置将无法加载，需迁移到新模型（如 Qwen2）。PR 正文已声明旧模型可在 vLLM <=0.23 中使用。 遗漏依赖：虽然有全面检查，但仍需确认是否有其他文件间接引用了被删除的符号（如 QWenBlock 被 qwen_vl.py 引用，已同步删除）。 测试覆盖：删除了相关测

试，可能降低对旧模型消失的验证，但旧模型已不再维护。

- 影响：用户影响：移除对旧模型的支持，用户需升级到新 Qwen 模型。文档需要同步更新（PR 已包含 documentation 标签）。 系统影响：减少代码库体积和维护负担，降低 CI 时间。 团队影响：减少维护责任，释放资源关注新架构。
- 风险标记：向后兼容性，跨版本迁移成本，遗漏引用

关联脉络

- PR #45127 [Model] Remove obsolete ERNIE models: 类似的旧模型移除 PR，评论中提到等待此 PR 合并以避免冲突。