PR #42616 完整报告

vllm-project/vllm

fix: propagate revision/code_revision pins to all artifact boundaries

合并时间：2026-05-15 17:31

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42616

执行摘要

- 一句话：修复 revision/code_revision 未传递到所有模型加载路径的问题
- 推荐动作：建议合并。这是正确性问题修复，变更小而专注，有测试覆盖（GGUF 路径），且所有改动的模式一致：在缺少 revision/code_revision 参数传递的地方补上。值得关注的设计决策是统一使用 model_config.revision 和 model_config.code_revision 作为唯一来源，避免后续新增加载路径再次遗漏。

功能与动机

用户通过 --revision 或 --code-revision 指定模型版本时，部分加载路径并未使用这些固定值，导致可能加载错误的模型版本（如未指定版本时可能加载非预期的最新版本）。PR 描述明确说明 This patch closes six pin-decay gaps so that every code, config, processor, and weight artifact loaded for a model resolves under the operator's configured revision。

实现拆解

1. GGUF 加载器 (vllm/model_executor/model_loader/gguf_loader.py): 在 _prepare_weights 方法中，针对 repo_id/filename.gguf 路径模式调用 hf_hub_download 时，新增传递 revision=model_config.revision 和 cache_dir=self.load_config.download_dir 参数，确保 GGUF 文件下载使用正确的版本。
2. Kimi Audio 模型 (vllm/model_executor/models/kimi_audio.py): 在 KimiAudioWhisperEncoder.__init__ 中调用 HFWhisperConfig.from_pretrained 时，新增传递 revision=vllm_config.model_config.revision；在 KimiAudioForCausalLM.__init__ 中定义 self.secondary_weights 时，将 revision=None 改为 revision=vllm_config.model_config.revision，确保 Whisper 配置和辅助权重文件都使用指定 revision。
3. 模型注册表 (vllm/model_executor/models/registry.py): 在 _try_resolve_transformers 方法中调用 try_get_class_from_dynamic_module 时，新增传递 code_revision=model_config.code_revision 参数，保证 remote code 的版本正确。
4. RoBERTa 模型 (vllm/model_executor/models/roberta.py): 在 BgeM3EmbeddingModel.__init__ 中定义 self.secondary_weights 时，将 revision=None 改为 revision=vllm_config.model_config.revision，确保辅助权重（如 sparse_linear.pt）也使用指定 revision。

5. Kimi K25 模型 (vllm/model_executor/models/kimi_k25.py): 在图像处理器初始化调用 cached_get_image_processor 时，新增传递 revision=self.ctx.model_config.revision 参数。
6. 测试 (tests/models/test_gguf_download.py): 更新单元测试 test_prepare_weights_repo_filename，验证 hf_hub_download 被调用时包含了 revision 和 cache_dir 参数。

关键文件：

- vllm/model_executor/model_loader/gguf_loader.py（模块 模型加载器；类别 source；类型 core-logic；符号 _prepare_weights）：修复 GGUF 文件下载时未传递 revision 的关键路径。
- vllm/model_executor/models/kimi_audio.py（模块 模型定义；类别 source；类型 core-logic；符号 KimiAudioWhisperEncoder.init, KimiAudioForCausalLM.init）：修复 Whisper 配置和辅助权重加载未传递 revision 的问题。
- vllm/model_executor/models/registry.py（模块 模型注册表；类别 source；类型 core-logic；符号 _try_resolve_transformers）：新增传递 code_revision 参数，确保 remote code 加载时使用正确的版本固定值。
- vllm/model_executor/models/roberta.py（模块 模型定义；类别 source；类型 core-logic；符号 BgeM3EmbeddingModel.init）：修复 BGE-M3 辅助权重加载未传递 revision 的问题。
- vllm/model_executor/models/kimi_k25.py（模块 模型定义；类别 source；类型 core-logic；符号 InputProcessor.init）：修复图像处理器加载未传递 revision 的问题。
- tests/models/test_gguf_download.py（模块 测试；类别 test；类型 test-coverage；符号 test_prepare_weights_repo_filename）：新增测试覆盖 GGUF 修改路径中的 revision 和 cache_dir 参数。

关键符号：GGUFModelLoader._prepare_weights, KimiAudioWhisperEncoder.init, KimiAudioForCausalLM.init, ModelRegistry._try_resolve_transformers, BgeM3EmbeddingModel.init, KimiK25InputProcessor.init

关键源码片段

vllm/model_executor/model_loader/gguf_loader.py

修复 GGUF 文件下载时未传递 revision 的关键路径。

```
def _prepare_weights(self, model_config: ModelConfig):
    model_name_or_path = model_config.model
    if os.path.isfile(model_name_or_path):
        return model_name_or_path
    # repo id/filename.gguf
    if "/" in model_name_or_path and model_name_or_path.endswith(".gguf"):
        repo_id, filename = model_name_or_path.rsplit("/", 1)
        # 修复：此前缺少 revision 参数，导致下载可能使用默认版本
        return hf_hub_download(
            repo_id=repo_id,
```

```
                filename=filename,
                revision=model_config.revision,
                cache_dir=self.load_config.download_dir,
            )
        # repo_id:quant_type
        elif "/" in model_name_or_path and ":" in model_name_or_path:
            repo_id, quant_type = model_name_or_path.rsplit(":", 1)
            return download_gguf(
                repo_id,
                quant_type,
                cache_dir=self.load_config.download_dir,
                revision=model_config.revision,
                ignore_patterns=self.load_config.ignore_patterns,
            )
        raise ValueError(
            f"Unrecognised GGUF reference: {model_name_or_path} "
            "(expected local file, <repo_id>/<filename>.gguf, "
            "or <repo_id>:<quant_type>)"
        )
```

评论区精华

没有实质性的 review 讨论。DarkLight1337 直接批准并感谢修复；gemini-code-assist[bot] 无反馈；claude[bot] 因来自 fork 跳过自动审核。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。每个修改都是单向参数传递增强，不会移除既有参数。潜在的回归在于如果某些加载路径依赖 revision=None 的行为（如拉取最新版本），强制传递用户指定的 revision 可能改变行为。但这是预期修正，与用户意图一致。测试覆盖了 GGUF 路径，其他变更依赖集成测试。
- 影响：影响中等偏正面。修复了当用户指定 --revision 或 --code-revision 时某些模型构件（GGUF 文件、Whisper 配置、辅助权重、remote code）可能加载错误版本的问题。影响范围限于使用这些参数的用户，特别是使用 Kimi-Audio、BGE-M3、GGUF 模型或自定义架构（依赖 remote code）的用户。
- 风险标记：缺少测试覆盖（部分路径），行为变更：formerly-default None 被覆盖

关联脉络

- PR #42444 [Model Runner V2][Bug Fix][DSV4] Ensure lazy attention state initializations happen during cudagraph capture: 同为修复参数传递遗漏类 bug，但影响模块不同。