PR #43321 完整报告

vllm-project/vllm

Correcting the mock classes for MM GC tests

合并时间：2026-05-22 15:21

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43321

执行摘要

- 一句话：修复 MM CUDA Graph 测试 mock 类缺失方法
- 推荐动作：可以快速合入。该 PR 是典型的测试配套修复，建议定期运行受影响的测试以确保 mock 与协议同步。关注 SupportsEncoderCudaGraph 协议的演进，避免再次出现类似不匹配。

功能与动机

修复因 PR #42224 引入的 postprocess_encoder_output 方法导致的多模态 CUDA Graph 测试失败。PR body 中明确列出了 8 个测试用例因 'SimpleMockViTModel' object has no attribute 'postprocess_encoder_output' 而失败，补丁后全部通过。

实现拆解

1. 在文件 tests/v1/cudagraph/test_encoder_cudagraph.py 顶部添加对 SupportsEncoderCudaGraph 的导入。
2. 让 _MockModel 类继承 SupportsEncoderCudaGraph，确保其符合协议要求。
3. 让 SimpleMockViTModel 类继承 SupportsEncoderCudaGraph，并移除原有的 supports_encoder_cudagraph = True 属性声明，因为协议继承已自动提供该标记。
4. 由于 SupportsEncoderCudaGraph 协议已经定义了 postprocess_encoder_output 方法的默认实现（作为抽象方法或带默认实现的方法，具体取决于协议定义），mock 类继承后自然获得了该方法，无需显式实现。

关键文件：

- tests/v1/cudagraph/test_encoder_cudagraph.py（模块 测试；类别 test；类型 test-coverage；符号 _MockModel, SimpleMockViTModel）：唯一修改的文件，通过让 mock 类继承 SupportsEncoderCudaGraph 协议来修复因协议新增方法导致的测试失败。

关键符号：_MockModel, SimpleMockViTModel

关键源码片段

tests/v1/cudagraph/test_encoder_cudagraph.py

唯一修改的文件，通过让 mock 类继承 SupportsEncoderCudaGraph 协议来修复因协议新增方法导致的测试失败。

```
# tests/v1/cudagraph/test_encoder_cudagraph.py

from typing import Any
import pytest
import torch
from vllm.model_executor.models.interfaces import SupportsEncoderCudaGraph # 新增导入
from vllm.platforms import current_platform
from vllm.v1.worker.encoder_cudagraph import EncoderCudaGraphManager
from vllm.v1.worker.encoder_cudagraph_defs import (
    EncoderCudaGraphCaptureInputs,
    EncoderCudaGraphConfig,
    EncoderCudaGraphReplayBuffers,
)

class _MockModel(SupportsEncoderCudaGraph): # 让 _MockModel 继承协议
    """Minimal mock implementing SupportsEncoderCudaGraph for __init__."""
    def __init__(self, min_budget: int = 4, max_budget: int = 128):
        self._min_budget = min_budget
        self._max_budget = max_budget
    def get_encoder_cudagraph_config(self) -> EncoderCudaGraphConfig:
        return EncoderCudaGraphConfig(
            modalities=["image"],
            input_key_by_modality={"image": "pixel_values"},
            buffer_keys=["dummy_buf"],
            out_hidden_size=32,
        )
    def get_encoder_cudagraph_budget_range(self, vllm_config):
        return (self._min_budget, self._max_budget)

class SimpleMockViTModel(torch.nn.Module, SupportsEncoderCudaGraph): # 让
SimpleMockViTModel 继承协议
    """Minimal ViT model for CUDA graph tests.
    Implements the SupportsEncoderCudaGraph protocol by providing
    all required methods. The forward pass projects patches and
    simulates spatial merge by averaging groups of m^2 patches.
    """
    def __init__(self):
        super().__init__()
        self.proj = torch.nn.Linear(_FLAT, _HIDDEN)
```

评论区精华

gemini-code-assist[bot] 评论指出 _MockModel 和 SimpleMockViTModel 缺少 get_max_frames_per_video 方法，可能影响类型检查。但 review 最终由 shen-shanshan 和 Isotr0py 批准并合并，表明该风险已被评估为可接受，或测试路径中不涉及 video 相关逻辑。

- Missing get_max_frames_per_video method in mocks (correctness): 未明确解决，但 review 最终获得批准。可能因为测试用例未涉及 video 路径，或者协议提供默认实现，因此当前不影响测试。

风险与影响

- 风险：风险较低。变更仅影响测试 mock 类，不涉及生产代码。如果 SupportsEncoderCudaGraph 协议将来新增抽象方法，这些 mock 类可能会再次失效。此外，get_max_frames_per_video 方法未实现，如果未来测试路径变化，可能暴露该缺失。
- 影响：影响范围仅限于 tests/v1/cudagraph/test_encoder_cudagraph.py 中的 8 个 CUDA Graph 测试用例。修复后这些测试恢复正常通过，确保多模态编码器 CUDA Graph 功能的回归测试覆盖。
- 风险标记：协议接口同步风险

关联脉络

- PR #42224 Add postprocess_encoder_output to SupportsEncoderCudaGraph: 本 PR 是为了适配 #42224 对协议接口的变更而进行的测试修复。