

PR #49339 完整报告

vllm-project/vllm

[CI] Fix and wire encoder/manager cudagraph unit tests

合并时间: 2026-07-22 03:41

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49339>

执行摘要

- 一句话: 修复未启用 CI 的 cudagraph 单元测试并接入 CI
- 推荐动作: 建议合并。这是一个典型的 CI 与测试修复 PR, 修复了测试遗漏和代码缺陷, 提升了测试覆盖和 CI 可靠性。

功能与动机

PR body 指出: `tests/v1/cudagraph/test_encoder_cudagraph.py` 和 `test_cudagraph_manager.py` 未被任何 Buildkite job 引用, 从未运行。且 encoder 测试因 `get_num_graphs_to_capture()` 读取 `self.config.enable_dual_path_graph` 但 `_make_manager_with_budgets` 未设置 `self.config` 导致 `AttributeError`。

实现拆解

1. 修复 encoder 测试中的 `AttributeError`: 在 `tests/v1/cudagraph/test_encoder_cudagraph.py` 的 `_make_manager_with_budgets` 函数中, 在通过 `object.__new__` 构造 `EncoderCudaGraphManager` 后, 新增了对 `mgr.config` 的赋值, 传入一个最小化的 `EncoderCudaGraphConfig` 对象, 确保 `get_num_graphs_to_capture()` 能正确读取 `enable_dual_path_graph` 等配置。
2. 将 encoder 测试接入 GPU CI: 在 `.buildkite/test_areas/cuda.yaml` 的 `cudagraph` job 的 `commands` 中添加 `pytest -v -s v1/cudagraph/test_encoder_cudagraph.py`, 并在 `source_file_dependencies` 中添加 `vllm/v1/worker/encoder_cudagraph.py` 和 `vllm/v1/worker/encoder_cudagraph_defs.py`, 确保源码变更会触发该测试。
3. 将 manager 测试接入 CPU CI: 在 `.buildkite/test_areas/misc.yaml` 的 `V1 Others (CPU)` job 的 `commands` 中添加 `pytest -v -s v1/cudagraph/test_cudagraph_manager.py`, 该测试标记为 `cpu_test`, 适合在 CPU 机器上运行。

关键文件:

- `tests/v1/cudagraph/test_encoder_cudagraph.py` (模块 `Encoder Cudagraph`; 类别 `test`; 类型 `test-coverage`; 符号 `_make_manager_with_budgets`): 修复了因缺少 `self.config` 导致的 `AttributeError`, 通过 `_make_manager_with_budgets` 函数中添加 `EncoderCudaGraphConfig` 赋值。
- `.buildkite/test_areas/cuda.yaml` (模块 `CI 配置`; 类别 `config`; 类型 `configuration`): 将 `encoder cudagraph` 测试添加到 GPU CI job 中, 并添加了对应的 `source file`

dependencies。

- .buildkite/test_areas/misc.yaml (模块 CI 配置; 类别 config; 类型 configuration) : 将 cudagraph manager 单元测试 (CPU 测试) 添加到 V1 Others (CPU) job 中。

关键符号: `_make_manager_with_budgets`

关键源码片段

tests/v1/cudagraph/test_encoder_cudagraph.py

修复了因缺少 `self.config` 导致的 `AttributeError`, 通过 `_make_manager_with_budgets` 函数中添加 `EncoderCudaGraphConfig` 赋值。

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

def _make_manager_with_budgets(budgets: list[int]) -> EncoderCudaGraphManager:
    """Create a minimal EncoderCudaGraphManager with only token_budgets set.

    Skips the parts of __init__ that require a real VllmConfig / model
    by patching the attributes directly after construction.
    """
    mgr = object.__new__(EncoderCudaGraphManager)
    mgr.token_budgets = sorted(budgets)
    mgr.max_batch_size = 16
    mgr.use_dp = False
    # 修复: 此前缺少 mgr.config 赋值, 导致 get_num_graphs_to_capture()
    # 读取 self.config.enable_dual_path_graph 时抛 AttributeError。
    # 现在直接设置一个最小化的 EncoderCudaGraphConfig 实例。
    mgr.config = EncoderCudaGraphConfig(
        modalities=["image"],
        buffer_keys=[],
        out_hidden_size=32,
    )
    mgr.budget_graphs = {"default": {}}
    mgr.graph_pool = None
    mgr.graph_hits = 0
    mgr.graph_misses = 0
    mgr.log_stats_interval = 100
    return mgr
```

评论区精华

本 PR 无 reviewer 评论。仅 claude[bot] 自动评论指出 PR 来自 fork 且自动 review 被禁用; hmellor 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及测试修复和 CI 配置，不影响任何生产代码。新增的 CI 依赖关系可能偶尔触发不必要的测试，但影响很小。
- 影响：
 - 用户：无直接影响。
 - 系统：CI 将自动运行 encoder cudagraph 和 manager 单元测试，防止回归。
 - 团队：需要维护 CI 配置中的依赖路径，但整体负担很小。
 - 风险标记：暂无

关联脉络

- PR #49302 [Bugfix] Fix DSA crash under breakable piecewise cudagraphs: 同样涉及 cudagraph 测试，修改了 tests/v1/cudagraph/test_breakable_cudagraph.py，与本 PR 的 cudagraph 测试文件属于同一测试目录。
- PR #48843 [BugFix] Set graph_pool_id before FULL CUDA graph capture in ModelRunner V2: 同样修复 cudagraph 相关问题，修改了 tests/v1/cudagraph/test_cudagraph_manager.py，与本 PR 接入了该测试的 CI 执行。