

PR #46780 完整报告

vllm-project/vllm

[ROCm] Fix AITER_UNIFIED_ATTN Dispatching After AITER Bump

合并时间: 2026-06-26 17:09

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46780>

执行摘要

- 一句话: 修复 AITER 升级后 unified attention 分发断言失败
- 推荐动作: 值得精读, 特别是对 ROCm 注意力后端的选择和测试模式感兴趣的同学。该 PR 体现了升级上游依赖 (AITER) 后如何快速定位并修复兼容性问题, 以及如何通过声明式约束和测试调整来保证正确性。此外, AMD CI 构建步骤硬失败的配置变更也值得关注。

功能与动机

AITER 升级到 v0.1.16.post2 后, 其 3D 内核增加了对 kv_cache_dtype 的断言, 仅支持 bf16 和 fp8, 导致使用 fp16 KV cache 的测试 (如 Whisper 测试) 触发 `AssertionError: kv_cache_dtype only supports BF16 (torch.bfloat16), FP8 (torch.float8_e4m3fnuz)`. 需要修复分发逻辑, 使该后端只接收支持的 dtype。

实现拆解

1. 声明后端支持的 dtype: 在 `vllm/v1/attention/backends/rocm_aiter_unified_attn.py` 中为 `RocmAiterUnifiedAttentionBackend` 添加了类变量 `supported_dtypes` (只包含 `torch.bfloat16`) 和 `supported_kv_cache_dtypes` (包含 `"auto"`、`"bfloat16"`、`"fp8"`、`"fp8_e4m3"`) ; 这一变更让上层的注意力后端选择和测试框架能够查询该后端的能力。
2. 修复后端选择测试: 在 `tests/v1/attention/test_rocm_attention_backends_selection.py` 中, 将 `test_standard_attention_backend_selection` 的 dtype 从固定的 `torch.float16` 改为根据 `selected_backend` 决定: 对于 `ROCM_AITER_UNIFIED_ATTN` 使用 `torch.bfloat16`, 否则使用 `torch.float16`。这样避免了选择测试本身使用不支持的 dtype 而失败。
3. 修复编译融合测试跳过逻辑: 在 `tests/compile/passes/test_fusion_attn.py` 的 `test_attention_quant_pattern` 中, 在获取 backend 类后检查 dtype 是否在 `supported_dtypes` 中, 若不在则跳过该测试, 并添加 TODO 注释等待 AITER 重新支持 fp16。
4. 限制单元测试 dtype: 在 `tests/kernels/attention/test_rocm_aiter_unified_attn.py` 中, 将 `DYPES` 列表从 `[torch.bfloat16, torch.float16]` 改为 `[torch.bfloat16]`, 并添加 TODO 注释。
5. 更新文档: 在 `docs/design/attention_backends.md` 中更新了 `ROCM_AITER_UNIFIED_ATTN` 行的 dtype 和 kv_cache_dtype 列。

6. CI 配置：在 `.buildkite/hardware_tests/amd.yaml` 中，将两个关键步骤（`ensure-ci-base-amd` 和 `image-build-amd`）的 `soft_fail` 设为 `false`，使构建失败立即硬失败，从而更可靠地捕获上游 CI 问题。

关键文件：

- `vllm/v1/attention/backends/rocm_aiter_unified_attn.py`（模块 注意力后端；类别 `source`；类型 `dependency-wiring`；符号 `supported_dtypes`, `supported_kv_cache_dtypes`）：添加了 `supported_dtypes` 和 `supported_kv_cache_dtypes` 类变量，是核心修复所在，使后端能力可查询。
- `tests/v1/attention/test_rocm_attention_backends_selection.py`（模块 测试；类别 `test`；类型 `test-coverage`；符号 `test_standard_attention_backend_selection`）：修复了后端选择测试中对 AITER 后端使用错误 `dtype` 的问题，确保测试通过。
- `tests/compile/passes/test_fusion_attn.py`（模块 测试；类别 `test`；类型 `test-coverage`；符号 `test_attention_quant_pattern`）：添加了根据 `supported_dtypes` 跳过测试的逻辑，避免使用不支持的 `dtype` 运行融合测试导致失败。
- `tests/kernels/attention/test_rocm_aiter_unified_attn.py`（模块 测试；类别 `test`；类型 `test-coverage`；符号 `DYPES`）：移除了 `DYPES` 中的 `torch.float16`，仅保留 `bf16`，避免触发内核断言。
- `.buildkite/hardware_tests/amd.yaml`（模块 CI 配置；类别 `test`；类型 `test-coverage`）：将两个关键 CI 步骤设为 `hard fail`，以确保上游问题能被及时发现。
- `docs/design/attention_backends.md`（模块 文档；类别 `docs`；类型 `documentation`）：更新了文档中 `ROCM_AITER_UNIFIED_ATTN` 行的 `dtype` 和 `kv_cache_dtype` 列，与代码一致。

关键符号：`test_standard_attention_backend_selection`, `test_attention_quant_pattern`

关键源码片段

`vllm/v1/attention/backends/rocm_aiter_unified_attn.py`

添加了 `supported_dtypes` 和 `supported_kv_cache_dtypes` 类变量，是核心修复所在，使后端能力可查询。

```
from typing import ClassVar
import torch
from vllm.config.cache import CacheDType
from vllm.v1.attention.backends.rocm_attn import RocmAttentionBackend
```

```
class RocmAiterUnifiedAttentionBackend(RocmAttentionBackend):
    # 声明该后端支持的输入 dtype（仅 bfloat16）
    supported_dtypes: ClassVar[list[torch.dtype]] = [torch.bfloat16]
    # 声明支持的 KV cache dtype（auto/bfloat16/fp8/fp8_e4m3）
    supported_kv_cache_dtypes: ClassVar[list[CacheDType]] = [
        "auto",
        "bfloat16",
        "fp8",
        "fp8_e4m3",
```

```
]
# ... (其余方法保持不变)
```

tests/v1/attention/test_rocm_attention_backends_selection.py

修复了后端选择测试中对 AITER 后端使用错误 dtype 的问题，确保测试通过。

```
# AITER unified attention kernel 只支持 BF16/FP8 KV cache
# 所以选择测试时必须使用 bf16
dtype = (
    torch.bfloat16
    if selected_backend == "ROCM_AITER_UNIFIED_ATTN"
    else torch.float16
)
attn_selector_config = AttentionSelectorConfig(
    head_size=128,
    dtype=dtype,
    # ...
)
```

评论区精华

仅有一条 maintainer 评论：AndreasKaratzas 批准该 PR，评论为 "Let's unblock upstream CI"，表明本 PR 旨在快速解除上游 CI 阻塞，功能改动聚焦于修复 dtype 分发。

- PR 快速合入以解除 CI 阻塞 (other): PR 被批准并合入。

风险与影响

- 风险：
 - 回归风险：低。变更主要是声明式约束 (supported_dtypes/kv_cache_dtypes)，并将测试 dtype 收紧到 bf16。但注意 supported_dtypes 当前只在测试中用于 skip 判断，实际 dispatching 应依赖现有入口逻辑 (get_attn_backend_cls)，后者需要确保不会向该后端传入不支持的 dtype—代码路径检查可能不足。
 - 功能限制：当前 ROCM_AITER_UNIFIED_ATTN 不再支持 fp16 输入，但该限制已在文档和测试中明确体现。若用户依赖 fp16 运行，将 fallback 到其他后端。
 - CI 影响：使 AMD CI 构建步骤硬失败，能尽早暴露问题，但可能增加临时失败。
- 影响：
 - 用户影响：ROCm 平台用户在使用 AITER unified attention 后端时，必须使用 bf16 或 fp8 KV cache；fp16 将导致回退或错误。
 - 系统影响：修复了 CI 中 Whisper 等测试的崩溃，AMD CI 构建步骤将严格失败。
 - 团队影响：低，变更范围清晰，影响面小。
 - 风险标记：仅通过测试验证分发逻辑，fp16 不再被支持

关联脉络

- PR #46780 [ROCm] Fix AITER_UNIFIED_ATTN Dispatching After AITER Bump: 本 PR 自身