

PR #51435 完整报告

vllm-project/vllm

[Bugfix][MM] Avoid device sync in FusedInputNorm initialization

合并时间: 2026-08-08 13:32

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51435>

执行摘要

- 一句话: FusedInputNorm 初始化去设备同步, 修复 AMD CI 多模态崩溃
- 推荐动作: 建议精读。这是一个小而典型的“默认设备上下文导致隐式同步”修复, 值得关注两个设计点: 一是在 CPU 上完成标量判定后再在目标设备上创建 buffer; 二是测试用 meta 设备模拟非 CPU 默认设备, 兼顾 CPU-only 分片。适合处理多模态初始化或设备同步问题的工程师阅读。

功能与动机

PR body 说明: PR #50411 将多模态归一化移上加速器, 同时把构建期的 torch.allclose 恒等检查也移到了加速器上, 引入了不必要的同步, 导致 Qwen2-VL sleep/wake 测试在 AMD CI (build 11806、11817) 失败。作者在讨论中表示, 该改动移除了 allclose 中导致崩溃的 GPU 到主机标量同步。因此目标是在不改变设备端预处理行为的前提下, 消除初始化期间的跨设备同步。

实现拆解

1. 修改 vllm/model_executor/models/vision.py 中 FusedInputNorm.init: 将恒等判定的中间张量显式创建在 device='cpu' (image_mean_cpu、image_std_cpu、weight_cpu、bias_cpu), 并基于 CPU 张量调用 torch.allclose 得出 is_identity。
2. 保持非恒等分支在默认设备上创建 image_mean_tensor 等张量并注册 weight / bias / running_mean / running_var 缓冲区; 恒等分支仍注册 None, 使前向语义完全不变。
3. 在 tests/models/multimodal/processing/test_qwen2_vl.py 新增 test_fused_input_norm_initialization_on_device, 用 monkeypatch 将 torch.allclose 替换为只接受 CPU 张量的 cpu_allclose, 并分别在 cuda (可用时) 或 meta 默认设备下验证恒等 / 非恒等两种参数组合。
4. 第二个提交把测试适配为 CPU-only 分片可运行 (meta 设备), 保证无 CUDA 环境下也能覆盖非 CPU 默认设备语义; 本地 MI300 上真实 Qwen2-VL sleep/wake 测试通过 (3 passed)。

关键文件:

- vllm/model_executor/models/vision.py (模块 视觉层; 类别 source; 类型 data-contract; 符号 FusedInputNorm): 核心修复: 将 FusedInputNorm 的恒等判定移到 CPU 常量张量, 消除构建期设备同步。

- tests/models/multimodal/processing/test_qwen2_vl.py (模块 多模态测试; 类别 test; 类型 test-coverage; 符号 test_fused_input_norm_initialization_on_device, cpu_allclose) : 新增针对 FusedInputNorm 初始化设备的测试, 用 monkeypatch 验证恒等判定只接触 CPU 张量, 并覆盖恒等 / 非恒等两条路径。

关键符号: FusedInputNorm.init, test_fused_input_norm_initialization_on_device, cpu_allclose

关键源码片段

vllm/model_executor/models/vision.py

核心修复: 将 FusedInputNorm 的恒等判定移到 CPU 常量张量, 消除构建期设备同步。

```
class FusedInputNorm(nn.Module):

    def __init__(
        self,
        image_mean: list[float],
        image_std: list[float],
        rescale_factor: float,
        channel: int = 3,
        dtype: torch.dtype = torch.float32,
    ) -> None:
        super().__init__()
        self.channel = channel

        # 模型构建期间可能把加速器设为 PyTorch 默认设备。
        # 在 CPU 上构造常量并做恒等判定, 避免 torch.allclose 在
        # 初始化阶段触发 GPU -> CPU 的隐式同步 (PR #51435 的修复点)。
        image_mean_cpu = torch.tensor(image_mean, dtype=dtype, device='cpu') * (
            1.0 / rescale_factor
        )
        image_std_cpu = torch.tensor(image_std, dtype=dtype, device='cpu') * (
            1.0 / rescale_factor
        )
        weight_cpu = 1.0 / image_std_cpu
        bias_cpu = -image_mean_cpu / image_std_cpu
        self.is_identity = bool(
            torch.allclose(weight_cpu, torch.ones_like(weight_cpu))
            and torch.allclose(bias_cpu, torch.zeros_like(bias_cpu))
        )

        if not self.is_identity:
            # 非恒等路径: 仍把归一化 buffer 放在调用方默认设备上,
            # 保留设备端预处理语义, 但 buffer 的构造仍可能在默认设备完成。
            image_mean_tensor = torch.tensor(image_mean, dtype=dtype) * (
                1.0 / rescale_factor
            )
            image_std_tensor = torch.tensor(image_std, dtype=dtype) * (
```

```

        1.0 / rescale_factor
    )
    weight = 1.0 / image_std_tensor
    bias = -image_mean_tensor / image_std_tensor
    self.register_buffer('weight', weight)
    self.register_buffer('bias', bias)
    self.register_buffer('running_mean', torch.zeros_like(image_mean_tensor))
    self.register_buffer('running_var', torch.ones_like(image_mean_tensor))
else:
    self.register_buffer('weight', None)
    self.register_buffer('bias', None)
    self.register_buffer('running_mean', None)
    self.register_buffer('running_var', None)

```

tests/models/multimodal/processing/test_qwen2_vl.py

新增针对 FusedInputNorm 初始化设备的测试，用 monkeypatch 验证恒等判定只接触 CPU 张量，并覆盖恒等 / 非恒等两条路径。

```

import pytest
import torch

from vllm.model_executor.models.vision import FusedInputNorm

@pytest.mark.parametrize(
    ('image_mean', 'image_std', 'rescale_factor', 'is_identity'),
    [
        ([0.0, 0.0, 0.0], [1.0, 1.0, 1.0], 1.0, True),
        ([0.5, 0.5, 0.5], [0.25, 0.25, 0.25], 1 / 255, False),
    ],
)
def test_fused_input_norm_initialization_on_device(
    monkeypatch: pytest.MonkeyPatch,
    image_mean: list[float],
    image_std: list[float],
    rescale_factor: float,
    is_identity: bool,
) -> None:
    """恒等判定不得同步默认设备：用包裹后的 allclose 断言输入在 CPU 上。"""
    original_allclose = torch.allclose

    def cpu_allclose(input: torch.Tensor, other: torch.Tensor, *args, **kwargs):
        assert input.device.type == 'cpu'
        assert other.device.type == 'cpu'
        return original_allclose(input, other, *args, **kwargs)

    monkeypatch.setattr(torch, 'allclose', cpu_allclose)
    # 优先使用真实加速器；CPU-only 分片用 meta 设备模拟非 CPU 默认设备。
    default_device = 'cuda' if torch.cuda.is_available() else 'meta'

```

```
with torch.device(default_device):
    input_norm = FusedInputNorm(image_mean, image_std, rescale_factor)
assert input_norm.is_identity is is_identity
if is_identity:
    assert input_norm.weight is None
    assert input_norm.bias is None
else:
    # 非恒等路径的 buffer 应位于调用方默认设备。
    assert input_norm.weight.device.type == default_device
    assert input_norm.bias.device.type == default_device
```

评论区精华

DarkLight1337 质疑非恒等时是否仍会同步，并解释直接初始化 GPU 张量仍会导致同步，除非先在 CPU 上初始化再显式 `non_blocking=True` 移到 GPU。noooooop 回应新实现比之前少一次同步，因为用于恒等判定的 `weight_cpu` 和 `bias_cpu` 都在 CPU 上。DarkLight1337 还追问这是否只是为了让测试通过，AndreasKaratzas 承认没有可复现的 `serve` 命令，但移除 `allclose` 的同步解决了 CI 崩溃。最后 DarkLight1337 提示启用该特性仍可能崩溃，该疑虑在批准前未完全闭环。

- `is_identity=False` 时是否仍存在同步 (design): 实现被接受，但非恒等路径在默认设备上创建 buffer 仍可能同步，讨论未完全消除该疑虑。
- 修复真实问题还是仅为通过测试 (question): 作者认为消除同步可解决 CI 崩溃，PR 最终被 noooooop 批准；真实 `serve` 场景能否复现仍未闭环。

风险与影响

- 风险：非恒等路径仍在默认设备上创建张量，该构造过程仍可能隐式同步，讨论中未彻底消除这项疑虑。改动只影响模型构建期，不改变前向计算数值，但所有走 `FusedInputNorm` 的多模态模型的加载阶段行为会统一变化。测试用 `meta` 设备模拟非 CPU 默认设备，不覆盖真实加速器的同步语义；真实行为只在本地 MI300 验证过。
- 影响：对用户而言，修复了 AMD CI 上 Qwen2-VL `sleep/wake` 测试失败，并可能消除模型加载阶段的隐式同步等待。对系统而言，`FusedInputNorm` 被多模态模型（如 Qwen2-VL）使用，构建期行为改变但前向结果不变。对团队而言，新增测试保证了恒等判定不再触碰设备张量，后续复用该模块时需维持 CPU 判定 + 默认设备建 buffer 的模式。
- 风险标记：非恒等路径同步疑虑未完全解决，`meta` 设备测试覆盖有限

关联脉络

- PR #50411（标题未提供）：PR body 明确指出该 PR 将多模态归一化移上加速器，导致 `FusedInputNorm` 构建期 `allclose` 在设备上执行并引入了本 PR 修复的同步问题。
- PR #51455 [Core] Make the GPU sync check thread-local and fix its suppressors: 同属 GPU 同步检查 / 消除主题，关注不必要的跨设备同步，与本 PR 在方向上一致。
- PR #51458 [Perf] Avoid some more unnecessary GPU<->CPU syncs: 同样以减少不必要 GPU-CPU 同步为目标，且涉及多模态模型文件，可互相参考。