

PR #29673 完整报告

sgl-project/sglang

[AMD][diffusion] fix: disable layernorm torch.compile decorator in eager mode on ROCm to avoid memory-access fault

合并时间: 2026-07-07 02:15

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29673>

执行摘要

- 一句话: 修复 ROCm 上 LayerNorm torch.compile 崩溃
- 推荐动作: 该 PR 值得快速合入, 属于必要 bugfix。建议关注后续是否有更通用的方案 (如全局 disable compile 开关), 但当前修复已足够。

功能与动机

在 8 卡 MI 系列 GPU 上以 eager 模式运行 Wan2.2-T2V-A14B 时, `ScaleResidual` 调制模块的 `forward_native` 因 `@torch.compile` 装饰器未被禁用, 触发 GPU 内存访问错误。该问题由 PR #25256 引入 (该 PR 仅对 NPU 禁用了 compile)。

实现拆解

1. 修改 `layernorm.py` 中的 `@torch.compile` 条件: 将 `forward_native` 方法上的 `torch.compile` 装饰器的 `disable` 参数从 `current_platform.is_npu()` 扩展为 `current_platform.is_npu() or current_platform.is_rocm()`。
2. 对同类方法应用相同修复: 在 `ScaleResidual` 类 (第 572 行)、`Fourier` 类 (第 740 行) 和 `TimestepEmbedBlock` 类 (第 837 行) 的 `forward_native` 方法上做了同样的条件扩展, 共计 4 处修改。
3. 保持行为一致: 所有修改仅影响编译行为, 不影响 eager 模式下的计算逻辑, 因此模型输出无变化。

关键文件:

- `python/sglang/multimodal_gen/runtime/layers/layernorm.py` (模块 扩散模型; 类别 source; 类型 core-logic; 符号 `forward_native`): 包含所有 `@torch.compile` 装饰器的修改, 是修复核心文件。

关键符号: `forward_native`

关键源码片段

`python/sglang/multimodal_gen/runtime/layers/layernorm.py`

包含所有 `@torch.compile` 装饰器的修改, 是修复核心文件。

```
# python/sglang/multimodal_gen/runtime/layers/layernorm.py
```

```

# 第 369 行 (LayerNorm 类) :
@torch.compile(
    backend="inductor",
    disable=current_platform.is_npu() or current_platform.is_rocm(),
)
def forward_native(self, x, residual=None):
    # ... eager 实现 ...

# 第 572 行 (ScaleResidual 类) :
@torch.compile(disable=current_platform.is_npu() or current_platform.is_rocm())
def forward_native(self, residual, x, gate, shift, scale):
    # ... eager 实现 ...

# 第 740 行 (Fourier 类) :
@torch.compile(disable=current_platform.is_npu() or current_platform.is_rocm())
def forward_native(self, x, shift, scale):
    # ... eager 实现 ...

# 第 837 行 (TimestepEmbedBlock 类) :
@torch.compile(disable=current_platform.is_npu() or current_platform.is_rocm())
def forward_native(self, x, scale, shift):
    # ... eager 实现 ...

```

评论区精华

无实质 review 讨论；mickqian 直接 Approved，HaiShaw 要求运行 pre-commit 以消除 lint 错误（已在第三次提交中修复）。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改 torch.compile 的禁用条件，不涉及核心逻辑变更。在 ROCm 上通过 eager 模式运行可避免内存错误，但可能因禁用 compile 而损失部分性能（若未来 ROCm 支持 compile 时需重新评估）。
- 影响：
 - 用户：在 ROCm 上使用 Wan2.2 或其他扩散模型的用户可正常启动，不再崩溃。
 - 系统：对 CUDA/XPU/MUSA/NPU 等平台无影响，仅 ROCm 行为改变。
 - 团队：无额外维护负担，修复简单且可逆。
 - 风险标记：仅影响 ROCm 平台

关联脉络

- PR #25256 previous change that introduced the regression (NPU-only guard): 该 PR 最初仅对 NPU 禁用了 torch.compile，导致 ROCm 上出现回归。