PR #31849 完整报告

sgl-project/sglang

fix(diffusion): keep fused qk-norm-rope out of dynamo tracing

合并时间：2026-07-28 19:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/31849

执行摘要

- 一句话：修复 diffusion dynamo tracing crash
- 推荐动作：值得立即合入。该 PR 精准定位了 @torch.compiler.assume_constant_result 与 SymInt 不兼容导致的问题，并且采用最小改动修复（增加守卫 + 改用静态 shape）。合入后可解除 torch.compile 与 diffusion 模型的兼容性阻塞。

功能与动机

启用 torch.compile（如 performance_mode=speed）后，所有 Qwen-Image 请求返回 HTTP 500，报错 torch._dynamo.exc.AsPythonConstantNotImplementedError: SymNodeVariable() is not a constant。原因是 can_use_fused_inplace_qknorm_rope 虽用 @torch.compiler.assume_constant_result 装饰，但调用时传入的 head_dim=img_query.shape[-1] 在 dynamo 下是 SymInt，无法被当作常量。

实现拆解

1. layernorm.py：在 apply_qk_norm_rope 函数的 fused 路径条件中增加 not torch.compiler.is_compiling() 守卫，当 dynamo 编译时跳过手写融合 kernel，让 unfused 路径被追踪并由 Inductor 自行融合。
2. qwen_image.py：将两处 apply_qk_norm_with_optional_rope 调用的 head_dim 参数从 img_query.shape[-1] / txt_query.shape[-1] 改为 self.head_dim，避免 dynamo 下产生 SymInt 参数。
3. 无测试配套变更。

关键文件：

- python/sglang/multimodal_gen/runtime/layers/layernorm.py（模块 归一化层；类别 source；类型 core-logic）：核心修复位置：在 fused 路径条件中增加 not torch.compiler.is_compiling() 守卫，避免 dynamo 追踪手写融合 kernel。
- python/sglang/multimodal_gen/runtime/models/dits/qwen_image.py（模块 扩散模型；类别 source；类型 data-contract）：调用方修复：将 head_dim 从运行时 shape 改为静态 self.head_dim，避免 dynamo 下 SymInt 参数导致 as_constant 失败。

关键符号：未识别

关键源码片段

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

核心修复位置：在 fused 路径条件中增加 not torch.compiler.is_compiling() 守卫，避免 dynamo 追踪手写融合 kernel。

```
# python/sglang/multimodal_gen/runtime/layers/layernorm.py
# 在 apply_qk_norm_rope 函数的 fused 路径条件判断中，增加编译期守卫：
# 当 torch.compiler.is_compiling() 为真时，跳过手写 fusion kernel，
# 让 unfused 路径被 dynamo 追踪并由 Inductor 自行融合。
if (
    fused_enabled
    and _is_cuda
    and not torch.compiler.is_compiling() # 新增：编译期跳过 fused kernel
    and allow_inplace
    and (q_eps == k_eps)
    and q.dtype in (torch.float16, torch.bfloat16)
    and q_norm.weight.dtype == q.dtype
    and k_norm.weight.dtype == k.dtype
    and q.is_contiguous()
    and k.is_contiguous()
    and can_use_fused_inplace_qknorm_rope(head_dim, rope_dim, is_neox, q.dtype)
):
    fused_inplace_qknorm_rope(...)
    return q, k
# 不满足 fused 条件时，走原有 unfused 路径
```

python/sglang/multimodal_gen/runtime/models/dits/qwen_image.py

调用方修复：将 head_dim 从运行时 shape 改为静态 self.head_dim，避免 dynamo 下 SymInt 参数导致 as_constant 失败。

```
# python/sglang/multimodal_gen/runtime/models/dits/qwen_image.py
# Qwen-Image attention forward 中的两处调用，将 head_dim 从运行时 shape
# ( 如 img_query.shape[-1]) 改为常量 self.head_dim，避免 dynamo 下的 SymInt 问题。
if self.qk_norm:
    img_query, img_key = apply_qk_norm_with_optional_rope(
        q=img_query,
        k=img_key,
        q_norm=self.norm_q,
        k_norm=self.norm_k,
        head_dim=self.head_dim, # 原为 img_query.shape[-1]
        cos_sin_cache=img_cache,
        is_neox=False,
        allow_inplace=True,
    )
    txt_query, txt_key = apply_qk_norm_with_optional_rope(
        q=txt_query,
        k=txt_key,
        q_norm=self.norm_added_q,
        k_norm=self.norm_added_k,
        head_dim=self.head_dim, # 原为 txt_query.shape[-1]
```

```
        cos_sin_cache=txt_cache,
        is_neox=False,
        allow_inplace=True,
    )
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：回归风险：低。仅当 torch.compiler.is_compiling() 为真时跳过 fused 路径，eager 模式行为不变。head_dim 改为 self.head_dim 是语义等价替换（因为 img_query.unflatten(-1, (self.local_num_heads, self.head_dim)) 已保证最后一维等于 self.head_dim）。 性能风险：compiling 路径下 fused kernel 不再使用，但 Inductor 仍可融合 unfused 操作，预期性能接近。 无测试覆盖：缺少对 compiling 场景的回归测试。
- 影响：用户影响：高——修复了torch.compile模式下Qwen-Image推理完全不可用的问题。 系统影响：低——仅影响 diffusion 模块中 Qwen-Image 模型的前向传播路径。 团队影响：低——3 行改动且逻辑清晰，易于 review。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #32616 [JIT] Restore the previous division behavior in per-token group quantization: 同样是 JIT / compile 场景下的 bugfix，体现了 SGLang 在处理 torch.compile 兼容性方面的持续投入。
- PR #32420 [diffusion] fix: preserve tensor stride when offloading rollout weights to pinned host memory: 同为 diffusion 模块 bugfix，修复了另一个令请求失败的严重问题，说明该模块正在稳定过程中。