PR #23266 完整报告

sgl-project/sglang

[NPU] [Bugfix] [Diffusion] Fixed gray images at the generation output

合并时间：2026-04-25 15:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23266

执行摘要

- 一句话：修复 NPU 扩散模型输出灰色图像的 RoPE 条件判断
- 推荐动作：值得一读，尤其是了解 NPU fallback 路径和 RoPE 实现的微妙之处。Reviewer 的建议展示了代码审查中对边界情况的敏感度。建议团队后续处理 reviewer 提出的两个潜在问题，提升 fallback 路径的鲁棒性。

功能与动机

引用 issue #23253 和 PR body：Qwen-image 和 FLUX 在 NPU 上输出灰色图像。原因是 PR #21633 引入的 if 条件导致 RoPE 对 interleaved 布局处理错误。

实现拆解

在 python/sglang/jit_kernel/diffusion/triton/npu_fallback.py 的 apply_rotary_embedding_native 函数中，原有条件 if cos.dim() == 3 and x.dim() == 3 and x.shape[1] < NPU_ROTARY_MUL_MAX_NUM_HEADS and x.shape[2] < NPU_ROTARY_MUL_MAX_HEAD_SIZE 会命中用于加速的 npu_rotary_mul 调用。但该加速不支持 interleaved 布局，导致 Qwen-image 等模型输出异常。

在原条件中新增 and not interleaved，使 interleaved=True 时绕过原生加速，进入手动 fallback 实现。

在 .github/workflows/pr-test-npu.yml 中将 python/sglang/jit_kernel/diffusion/triton/npu_fallback.py 添加到 multimodal_gen 路径过滤器中，确保 NPU CI 在修改此文件时被触发。

关键文件：

- python/sglang/jit_kernel/diffusion/triton/npu_fallback.py（模块 NPU 回退；类别 source；类型 core-logic）：核心修复：在 RoPE fallback 条件中加入 and not interleaved，避免使用 NPU 原生旋转乘法导致灰度图
- .github/workflows/pr-test-npu.yml（模块 CI 配置；类别 infra；类型 infrastructure）：将 nup_fallback.py 加入 CI 变更过滤器，确保 NPU CI 在修改此文件时自动触发

关键符号：apply_rotary_embedding_native

关键源码片段

python/sglang/jit_kernel/diffusion/triton/npu_fallback.py

核心修复：在 RoPE fallback 条件中加入 and not interleaved，避免使用 NPU 原生旋转乘法导致灰度图

```
def apply_rotary_embedding_native(
    x: torch.Tensor, cos: torch.Tensor, sin: torch.Tensor, interleaved: bool = False
) -> torch.Tensor:
    # 1. 确保 cos/sin 维度和类型正确
    cos = cos.unsqueeze(-2).to(x.dtype)
    sin = sin.unsqueeze(-2).to(x.dtype)

    # 2. 仅当满足形状约束且非 interleaved 时使用 NPU 原生旋转乘法
    # 原生 npu_rotary_mul 不支持 interleaved 布局，添加 not interleaved 条件修复灰度图
    if (
        cos.dim() == 3
        and x.dim() == 3
        and x.shape[1] < NPU_ROTARY_MUL_MAX_NUM_HEADS
        and x.shape[2] < NPU_ROTARY_MUL_MAX_HEAD_SIZE
        and not interleaved # <--- 本次修复新增条件
    ):
        # 如果 cos 是半尺寸，复制拼接为全尺寸
        if cos.size(-1) * 2 == x.size(-1):
            cos = torch.cat([cos, cos], dim=-1)
            sin = torch.cat([sin, sin], dim=-1)
        cos = cos.unsqueeze(0)
        sin = sin.unsqueeze(0)
        x = x.unsqueeze(0)
        x_embed = torch_npu.npu_rotary_mul(x, cos, sin)
        x_embed = x_embed.squeeze(0)
        return x_embed

    # 3. 手动 fallback（注意：当前实现硬编码为 interleaved 布局）
    # 潜在风险：当由其他原因 fallback 且 interleaved=False 时，结果可能错误
    x1 = x[..., ::2]
    x2 = x[..., 1::2]
    o1 = x1 * cos - x2 * sin
    o2 = x2 * cos + x1 * sin
    return torch.stack((o1, o2), dim=-1).flatten(-2)
```

评论区精华

Reviewer @gemini-code-assist[bot] 提出了两个潜在问题：

1. 手动 fallback 硬编码 interleaved 布局：手动实现使用 x[..., ::2] 和 x[..., 1::2] 拆分，这是 interleaved 布局的特有操作。如果因其它条件（如形状不满足）导致非 interleaved 数据也 fallback 到此路径，输出将不正确。
2. cos/sin 形状不匹配：当 interleaved=True 走手动 fallback 时，cos 和 sin 可能是全尺寸（cos.shape[-1] == x.shape[-1]），但手动实现期望 cos 只包含半尺寸（因为只在半维度上应用）。若 cos 未预先切片，可能导致形状不匹配。

两个问题均未在本 PR 中解决，但 PR 获得了批准（@ping1jing2、@Makcum888e）。团队可能认为当前改动足以解决紧急 bug，剩余问题可在后续迭代中修复。

- Fallback 路径硬编码 interleaved 布局 (correctness): 未在 PR 中解决，但 PR 已被批准合并；潜在风险遗留
- interleaved 回退时 cos/sin 形状适配 (correctness): 未处理，但可能暂时没有触发 bug；建议添加切片逻辑

风险与影响

- 风险：
  - 正确性风险：手动 fallback 路径硬编码为 interleaved 布局，当因其他条件（如维度不足）导致非 interleaved 数据进入 fallback 时，会生成错误旋转嵌入。影响范围：所有通过 apply_rotary_embedding_native 且设置 interleaved=False 但形状不满足原生加速条件的模型。
  - 兼容性风险：cos/sin 全尺寸与半尺寸假设不一致，可能导致维度错误。
  - 回归风险：低，改动仅为附加条件，不影响现有正确路径。
  - 性能风险：无，仅改变了路径选择。
  - 覆盖风险：未添加单元测试覆盖 interleaved/non-interleaved 分支。
- 影响：
  - 用户影响：修复了 NPU 上 Qwen-image 和 FLUX 等扩散模型输出全灰图像的问题，恢复模型可用性。
  - 系统影响：仅修改 NPU fallback 路径，不影响 CUDA 或其他平台。CI 配置变更确保未来相关修改能被自动测试。
  - 团队影响：低风险、易回滚的修复。但遗留的 review 建议可能引发后续 fixup PR。
  - 风险标记：手动 fallback 硬编码假设，cos/sin 切片兼容性，未解决 review 建议

关联脉络

- PR #21633 [NPU] Add condition for native rotary embedding: 引入了导致灰度图的条件判断，是本 PR 修复的直接原因