PR #23538 完整报告

sgl-project/sglang

[NPU] Fix Z-Image negative-branch rotary embeddings for CFG

合并时间：2026-05-03 21:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23538

执行摘要

- 一句话：修复 Z-Image 负提示旋转嵌入使用正提示长度的 bug
- 推荐动作：该 PR 值得审阅以理解扩散模型中 CFG 分支处理的常见陷阱；设计简单明了，适合作为 bugfix 范例。

功能与动机

Z-Image 在使用 CFG 生成图像时，负分支的旋转位置编码形状错误（32 vs 192），导致 Tensor 尺寸不匹配的运行时错误。该 Bug 由 PR body 中提供的堆栈跟踪和复现步骤明确报告。

实现拆解

1. 修改 prepare_neg_cond_kwargs 方法（python/sglang/multimodal_gen/configs/pipeline_configs/zimage.py，第 363-383 行）：新增 prompt_embeds 变量，优先使用 batch.negative_prompt_embeds[0]（若存在），否则回退到 batch.prompt_embeds[0]。将 get_freqs_cis 的第一个参数从此前的 batch.prompt_embeds[0] 替换为 prompt_embeds，确保负分支使用正确的嵌入长度。
2. 新增单元测试（python/sglang/multimodal_gen/test/unit/test_zimage_pipeline_config.py，全文件）：添加 TestZImagePipelineConfig.test_zimage_negative_prompt_rotary_embeddings_use_negative_prompt_len 方法，模拟不同正 / 负序列长度（19 vs 45），断言 prepare_neg_cond_kwargs 返回的 freqs_cis 中位置 ID 的形状与负提示序列长度对齐，验证修复正确性。

关键文件：

- python/sglang/multimodal_gen/configs/pipeline_configs/zimage.py（模块 扩散配置；类别 source；类型 core-logic；符号 prepare_neg_cond_kwargs）：核心修复：修改 prepare_neg_cond_kwargs 以使用负提示嵌入的长度构建 RoPE。
- python/sglang/multimodal_gen/test/unit/test_zimage_pipeline_config.py（模块 测试；类别 test；类型 test-coverage；符号 TestZImagePipelineConfig, test_zimage_negative_prompt_rotary_embeddings_use_negative_prompt_len）：新增单元测试验证修复，确保负分支使用负提示长度。

关键符号：prepare_neg_cond_kwargs, get_freqs_cis, test_zimage_negative_prompt_rotary_embeddings_use_negative_prompt_len

关键源码片段

python/sglang/multimodal_gen/configs/pipeline_configs/zimage.py

核心修复：修改 prepare_neg_cond_kwargs 以使用负提示嵌入的长度构建 RoPE。

```
# python/sglang/multimodal_gen/configs/pipeline_configs/zimage.py
class ZImagePipelineConfig:
    def prepare_neg_cond_kwargs(self, batch, device, rotary_emb, dtype):
        # 修复：使用负提示嵌入（如果存在），否则回退到正提示嵌入
        prompt_embeds = (
            batch.negative_prompt_embeds[0]
            if batch.negative_prompt_embeds is not None
            else batch.prompt_embeds[0]
        )
        return {
            "freqs_cis": self.get_freqs_cis(
                prompt_embeds, # 之前这里错误地使用了 batch.prompt_embeds[0]
                batch.width,
                batch.height,
                device,
                rotary_emb,
                batch,
            ),
            "image_seq_len_target": (
                self._get_zimage_sp_plan(batch)["img_seq_target"]
                if get_sp_world_size() > 1
                else None
            ),
        }
```

python/sglang/multimodal_gen/test/unit/test_zimage_pipeline_config.py

新增单元测试验证修复，确保负分支使用负提示长度。

```
# python/sglang/multimodal_gen/test/unit/test_zimage_pipeline_config.py
import unittest
from types import SimpleNamespace
from unittest.mock import patch

import torch

from sglang.multimodal_gen.configs.pipeline_configs.zimage import ZImagePipelineConfig


class TestZImagePipelineConfig(unittest.TestCase):
    @patch("sglang.multimodal_gen.configs.pipeline_configs.zimage.get_sp_world_size")
    def test_zimage_negative_prompt_rotary_embeddings_use_negative_prompt_len(
        self, mock_get_sp_world_size
    ) -> None:
        """Negative CFG branch should build RoPE positions from negative prompt embeds."""
```

```
        mock_get_sp_world_size.return_value = 1

        config = ZImagePipelineConfig()
        pos_seq_len = 19
        neg_seq_len = 45
        batch = SimpleNamespace(
            prompt_embeds=[torch.ones(pos_seq_len, 2560)],
            negative_prompt_embeds=[torch.ones(neg_seq_len, 2560)],
            height=16,
            width=16,
        )

        def rotary_emb(pos_ids):
            return pos_ids

        neg_kwargs = config.prepare_neg_cond_kwargs(
            batch=batch,
            device=torch.device("cpu"),
            rotary_emb=rotary_emb,
            dtype=torch.float32,
        )

        cap_pos_ids, image_pos_ids = neg_kwargs["freqs_cis"]
        neg_cap_padded_len = 64
        # 断言：caption 位置 ID 的形状应为 (64, 3)，基于负提示填充长度
        self.assertEqual(cap_pos_ids.shape, (neg_cap_padded_len, 3))
        # 断言：第一个图像位置 ID 正确反映了填充偏移
        self.assertEqual(image_pos_ids[0].tolist(), [neg_cap_padded_len + 1, 0, 0])


if __name__ == "__main__":
    unittest.main()
```

评论区精华

审查者 OrangeRedeng 要求添加 CI 测试以避免未来回归，贡献者 gxxx-hum 同意并提交了测试。合并者 ping1jing2 指出 GPU CI 出现另一个错误（由 #23625 引起）并确认 NPU CI 正常后合并。

- 添加 Z-Image CI 测试 (testing): 但该 PR 仅添加了单元测试，未集成到 CI 流水线；测试在 Python 端已覆盖。
- GPU CI 错误 (other): 确认无关后合并。

风险与影响

- 风险：风险极低：变更仅影响 Z-Image 模型的负分支 RoPE 构造，且逻辑简单（首选负提示嵌入，降级到正提示）。单元测试覆盖了核心场景，不会影响其他模型或正常分支。GPU CI 的失败与此次 PR 无关。

- 影响：影响范围仅限于使用 Z-Image 模型且启用 CFG（Classifier-Free Guidance）的用户。修复后，具有负提示的生成将正确工作，消除尺寸不匹配错误。无统计效果或兼容性问题。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR