# PR #25305 完整报告

sgl-project/sglang

[diffusion] Fix Z-Image Cache-DiT sequence-parallel override

合并时间：2026-05-15 13:25

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25305

## 执行摘要

- 一句话：修复 Cache-DiT 下 Z-Image 的 sequence parallel 覆盖问题
- 推荐动作：此 PR 值得精读，尤其是其修复模式——通过参数传递替代直接属性修改，是一种更稳健的设计。对于涉及模型包装和参数覆盖的场景有参考价值。

## 功能与动机

关联 Issue #25254 报告了当启用 Cache-DiT（SGLANG_CACHE_DIT_ENABLED=true）时，Z-Image 模型在推理时崩溃，返回 Internal Server Error。 PR body 描述了根本原因是 Cache-DiT 包装了 transformer blocks，导致 layer.attention 不再是直接的 ZImageAttention 实例，原先直接修改 layer.attention.attn.skip_sequence_parallel 的代码失效。

## 实现拆解

1. 在 ZImageAttention.forward 中新增参数：在文件 python/sglang/multimodal_gen/runtime/models/dits/zimage.py 的 ZImageAttention.forward 方法签名中添加 skip_sequence_parallel_override: bool = False 参数。
2. 在 ZImageAttention.forward 内部传递参数：当调用 self.attn(...) 时，将新增的 skip_sequence_parallel_override 参数传递给 USPAttention 的 forward 方法。
3. 在 ZImageTransformerBlock.forward 中新增并传递参数：在 ZImageTransformerBlock.forward 方法签名中添加 skip_sequence_parallel_override 参数，并在调用 self.attention(...) 时传递该参数。
4. 在 ZImageFinalLayer.forward 中新增并传递参数：类似地，在 ZImageFinalLayer.forward 方法签名中添加并传递该参数。
5. 修改主循环调用点：在 ZImageModel.forward 中，移除原先在每个 step 中直接设置 layer.attention.attn.skip_sequence_parallel = use_full_unified_sequence 的循环，改为在调用 layer(...) 时通过 skip_sequence_parallel_override=use_full_unified_sequence 参数传递。
6. 无测试文件变更：本次修改未包含测试文件，仅源码变更。

关键文件：

- python/sglang/multimodal_gen/runtime/models/dits/zimage.py（模块 扩散模型；类别 source；类型 core-logic；符号 ZImageAttention.forward, ZImageTransformerBlock.forward, ZImageFinalLayer.forward, ZImageModel.forward）：包含所有变更：在 ZImageAttention、ZImageTransformerBlock、ZImageFinalLayer 的 forward 方法中新增 skip_sequence_parallel_override 参数传递，并修改主循环调用方式。

关键符号：ZImageAttention.forward, ZImageTransformerBlock.forward, ZImageFinalLayer.forward, ZImageModel.forward

## 关键源码片段

### python/sglang/multimodal_gen/runtime/models/dits/zimage.py

包含所有变更：在 ZImageAttention、ZImageTransformerBlock、ZImageFinalLayer 的 forward 方法中新增 skip_sequence_parallel_override 参数传递，并修改主循环调用方式。

```
# python/sglang/multimodal_gen/runtime/models/dits/zimage.py
# 修改 ZImageAttention.forward：新增 skip_sequence_parallel_override 参数并传递给 self.attn
class ZImageAttention(nn.Module):
    def forward(
        self,
        hidden_states: torch.Tensor,
        freqs_cis: Optional[Tuple[torch.Tensor, torch.Tensor]] = None,
        num_replicated_prefix: int = 0,
        num_replicated_suffix: int = 0,
        skip_sequence_parallel_override: bool = False, # 新增参数，用于替代直接修改 layer.
        attention.attn.skip_sequence_parallel
    ):
        # ... (q/k/v 计算和形状变换 ) ...
        if num_replicated_suffix > 0:
            # ... (ulysses_attn 分支 ) ...
        else:
            hidden_states = self.attn(
                q,
                k,
                v,
                num_replicated_prefix=num_replicated_prefix,
                num_replicated_suffix=num_replicated_suffix,
                skip_sequence_parallel_override=skip_sequence_parallel_override, # 传递参数
            )
        # ...

# 修改 ZImageTransformerBlock.forward：传递 skip_sequence_parallel_override
class ZImageTransformerBlock(nn.Module):
    def forward(
        self,
        hidden_states: torch.Tensor,
        freqs_cis: Optional[Tuple[torch.Tensor, torch.Tensor]] = None,
        adaln_input: Optional[torch.Tensor] = None,
        num_replicated_prefix: int = 0,
```

```
        num_replicated_suffix: int = 0,
        skip_sequence_parallel_override: bool = False, # 新增参数
    ):
        # ...
        attn_out = self.attention(
            hidden_states,
            freqs_cis=freqs_cis,
            num_replicated_prefix=num_replicated_prefix,
            num_replicated_suffix=num_replicated_suffix,
            skip_sequence_parallel_override=skip_sequence_parallel_override, # 传递
        )
        # ...

# 修改 ZImageModel.forward：移除直接属性修改，改用参数传递
class ZImageModel(nn.Module):
    def forward(self, ...):
        # ...
        for layer in self.layers:
            unified = layer(
                unified,
                unified_freqs_cis,
                adaln_input,
                num_replicated_suffix=num_replicated_suffix,
                skip_sequence_parallel_override=use_full_unified_sequence, # 通过参数传递
            )
        # ...
```

## 评论区精华

PR 无 review 评论讨论，由 mickqian 直接批准。

- 暂无高价值评论线程

## 风险与影响

- 风险：
  1. 回归风险低：修改集中于单个文件 zimage.py，且改动量小（+7/-2），逻辑清晰，不易引入回归。
  2. 影响范围受限：仅影响启用 Cache-DiT 的 Z-Image 模型推理路径。
  3. 缺少测试覆盖：本次修改未包含对应测试用例，未来重构或修改可能缺乏保护。
- 影响：
  - 用户影响：修复了 Z-Image 模型在 Cache-DiT 下的崩溃问题，用户可正常使用图像生成功能。
  - 系统影响：无性能影响，改动仅在推理路径中添加了一个布尔参数传递。
  - 团队影响：为后续 Cache-DiT 和其他 diffusion 模型的集成提供了更安全的参数传递模式。
  - 风险标记：缺少测试覆盖

# 关联脉络

- 暂无明显关联 PR