PR #25592 完整报告

sgl-project/sglang

[Diffusion] [NPU] Fix HunyuanVideo crash on NPU

合并时间：2026-05-19 17:40

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25592

执行摘要

- 一句话：修复 NPU 上 Tensor.view 非连续张量崩溃
- 推荐动作：这是一个最小化、安全的 bugfix，值得直接合入。无需精读，但可作为 NPU 兼容性修复的示例参考。

功能与动机

HunyuanVideo 在 Ascend NPU 上执行时崩溃，错误日志显示 view() 调用因张量内存不连续而失败。PR body 说明 NPU 的 Torch SDPA 后端返回非连续张量，需改用 reshape()。

实现拆解

1. 定位崩溃点：在 python/sglang/multimodal_gen/runtime/models/dits/hunyuanvideo.py 的 MMDoubleStreamBlock.forward() 中，img_attn.view(batch_size, image_seq_len, -1) 因张量不连续而抛出 RuntimeError。
2. 替换方法：将 view() 改为 reshape()，二者在连续张量上行为一致，但 reshape() 能自动处理非连续情况。
3. 无其他文件修改：仅此一行变更，不影响其他逻辑。

关键文件：

- python/sglang/multimodal_gen/runtime/models/dits/hunyuanvideo.py（模块 扩散模型；类别 source；类型 data-contract）：唯一修改的文件，将 view() 替换为 reshape() 以兼容 NPU 上的非连续张量。

关键符号：未识别

关键源码片段

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

唯一修改的文件，将 view() 替换为 reshape() 以兼容 NPU 上的非连续张量。

```
# 在 MMDoubleStreamBlock 的 forward 方法中，原代码使用 .view()
# .view() 要求张量内存连续，而 NPU 的 SDPA 后端返回非连续张量
# 改用 .reshape() 自动处理不连续情况，行为在连续时与 .view() 一致
img_attn_out, _ = self.img_attn_proj(
    img_attn.reshape(batch_size, image_seq_len, -1) # 原为 .view(...)
```

）

评论区精华

无实质性审查讨论。机器人审查人（gemini-code-assist[bot]）确认变更合理。负责人 ping1jing2 直接批准。作者在合并后补充注释，解释 NPU 默认 SDPA 后端与 CUDA Flash Attention 的行为差异。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。reshape() 在连续张量上行为与 view() 相同，不会引入性能或语义回归。但需确认非 NPU 后端（CUDA）的 attn 输出是否可能也为非连续——从 error log 看，CUDA 使用 Flash Attention 返回连续张量，因此无影响。
- 影响：影响范围局限于 HunyuanVideo 模型在 NPU 上的推理通路。修复后 NPU 用户可正常执行生成。对其他硬件和模型无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR