PR #22266 完整报告

sgl-project/sglang

[NPU] fix qwen3.5 video processor

合并时间：2026-04-08 21:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22266

执行摘要

本 PR 修复了 NPU 平台上 Qwen3.5 视频处理器因超过 8 维张量置换操作导致的推理失败问题。通过重构预处理逻辑、提取公共函数并修正关键 bug，确保了视频输入在 NPU 上的正确处理，提升了多模态功能的稳定性和跨硬件兼容性。

功能与动机

动机源于 NPU 硬件对张量维度数的限制：Qwen3VLVideoProcessor 中的 permute 操作超过 8 维，在 NPU 上不受支持，导致视频推理失败。PR body 引用 issue 或相关 PR #20189，明确指出目标是“避免处理超过 8 维的数据”。此修复使 NPU 用户能够正常使用视频输入进行多模态推理。

实现拆解

主要改动在文件 python/sglang/srt/hardware_backend/npu/modules/qwen_vl_processor.py：

- 新增函数 transform_patches_to_flatten：将补丁张量通过视图和置换重构，避免高维操作。python def transform_patches_to_flatten(patches, batch_size, grid_t, ...): patches = patches.view(...) patches = patches.permute(...) return flatten_patches
- 修改 _preprocess 函数：调用新函数替换原有高维 permute 逻辑，简化代码结构。
- 添加 npu_wrapper_video_preprocess 包装器：覆盖视频预处理方法，处理 size 参数访问（如 size['shortest_edge']）并修正维度顺序。

评论区精华

Review 讨论聚焦于代码正确性和设计优化：

- gemini-code-assist[bot] 指出置换顺序错误：> “The permutation order (0, 1, 4, 3, 2, 5, 6) swaps the channel and temporal_patch_size dimensions... This swap will lead to incorrect model inputs.” 建议改为 (0, 1, 4, 2, 3, 5, 6) 以确保输入正确。
- gemini-code-assist[bot] 强调 size 参数处理：> “size is a SizeDict... should be accessed using bracket notation... handle the case where it is None.” 避免运行时错误。
- xiaobaicxy 建议代码重构：> “Please extract a common function” 以提升复用性，提交历史显示作者响应并实施。

风险与影响

- 技术风险：置换逻辑修正可能引入新 bug，需确保与原始实现一致性；size 参数处理不当可能导致 AttributeError；变更集中于 NPU 后端，可能影响其他硬件兼容性；PR 未添加单元测试，存在覆盖不足风险。
- 影响分析：直接影响 NPU 平台的 Qwen3.5 视频模型推理，修复了执行失败问题，用户现可正常使用视频输入。对系统，增强了 NPU 后端稳定性和多模态功能完整性，影响范围限于特定硬件模块。

关联脉络

从历史 PR 看，本 PR 与 #21692（NPU 上 Qwen3.5 量化修复）相关，共同构成 NPU 平台对 Qwen3.5 模型的全面支持。近期 PR 如 #22292 和 #22346 聚焦 CI 优化，而本 PR 延续了硬件适配趋势，揭示了 sglang 项目在跨平台兼容性（如 NPU、AMD）和多模态扩展上的持续演进。