PR #6993 完整报告

PaddlePaddle/FastDeploy

[XPU] Refactor pre process

合并时间：2026-04-01 20:29

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/6993

执行摘要

本次 PR 重构了 FastDeploy 中 XPU 平台的前处理逻辑，通过新增 speculate_pre_process 和 unified_update_model_status 算子，并移除多个算子中的 cum_offsets 参数，统一了推测解码的数据流。变更涉及 36 个文件，影响核心处理路径，旨在提升代码清晰度和处理效率，但存在回归风险和测试覆盖不足的问题。

功能与动机

为什么做：根据 PR body，动机是“前处理优化统一”，引用相关 PR #6358 和 #6501，旨在解决 XPU 上 MTP（混合令牌并行）推测解码场景的前处理逻辑碎片化问题，优化数据传递并修复解码异常。PR body 中展示了基准测试结果，表明在 21B A3B 单卡上解码结果发现明显异常，需通过重构来改进。

实现拆解

按模块拆解改动：

1. 新增算子层：在 custom_ops/xpu_ops/src/ops/mtp/ 下添加：
   - speculate_preprocess.cc：处理输入 ID、序列长度和草稿令牌，输出去除填充的 ID、批次映射、累积序列长度等结构化数据。
   - unified_update_model_status.cc：统一更新模型状态，包括序列长度、停止标志等，简化状态管理。
2. 算子参数重构：修改多个现有 XPU 算子（如 adjust_batch.cc、block_attn.cc、gather_next_token.cc），移除 cum_offsets 参数，改用 cu_seqlens_q_output 和 batch_id_per_token_output 等新参数，示例如下：cpp // 修改前 std::vector<paddle::Tensor> AdjustBatchKernel(const paddle::Tensor &x, const paddle::Tensor &cum_offsets, // 被移除 ...); // 修改后 std::vector<paddle::Tensor> AdjustBatchKernel(const paddle::Tensor &x, ...); // cum_offsets 已移除
3. Python 层集成：更新 fastdeploy/model_executor/xpu_pre_and_post_process.py 中的 xpu_pre_process 函数，调用新算子并调整数据结构传递，移除对 cum_offsets 的依赖。
4. 测试补充：新增 custom_ops/xpu_ops/test/test_speculate_pre_process.py 和 test_unified_update_model_status.py 单元测试文件，但 Codecov 报告显示 patch 覆盖率仅 50%。

评论区精华

Review 讨论摘要：评论中仅有两个 reviewer（hong19860320 和 freeliuzc）给出了 LGTM 并标记为 DISMISSED，没有具体技术交锋或争议点。这表明变更被快速接受，但可能缺乏对设计权衡的深度讨论。

风险与影响

具体风险：

- 回归风险：移除 cum_offsets 参数可能破坏现有 GPU/XPU 混合部署的兼容性，需验证跨平台一致性。
- 性能风险：新算子可能引入额外计算开销，影响实时解码性能，尤其是在高负载场景。
- 测试覆盖不足：Patch 覆盖率仅 50%，缺少边界情况测试，可能隐藏逻辑错误。

影响评估：

- 对用户：透明变更，但需确保推理结果无退化；基准测试显示解码异常，重构后应改善。
- 对系统：简化数据流，提升可维护性；但变更涉及核心路径，需谨慎部署。
- 对团队：工程师需学习新的数据结构，可能影响后续开发节奏。

关联脉络

与历史 PR 的关系：本 PR 直接引用 PR #6358（refactor MTP pre_process）和 #6501（speculate_pre_process），是前处理优化链条的一部分。近期历史 PR 中，如 #7001（Support mtp overlap schedule）和 #7107（PD Disaggregation）也涉及推测解码和调度优化，表明仓库正持续演进 MTP 和推测解码功能，本 PR 为这一方向的基础设施重构。