PR #7300 完整报告

PaddlePaddle/FastDeploy

[BugFix] Fix mtp empty run issue in overlap schedule and EP model

合并时间：2026-04-10 18:29

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7300

执行摘要

本次 PR 修复了 FastDeploy 中 overlap 调度模式下 MTP（推测解码）空输入未执行的问题，通过添加缺失的条件判断和函数调用，确保 EP 模型在空输入场景下的行为与正常调度模式一致。变更范围小，风险较低，但需注意测试覆盖不足。

功能与动机

根据 review 中的 AI 代码审查，问题背景是：在 overlap schedule 模式下，当 real_bsz == 0 或 model_output is None 时，系统未调用 MTP 的空输入处理函数，导致 EP 模型出现空运行问题。这与 execute_model_normal 函数中的逻辑不一致，可能影响系统可靠性。

实现拆解

修改仅涉及 fastdeploy/worker/gpu_model_runner.py 文件中的 execute_model_overlap 函数。在函数的 else 分支（处理空输入场景）中添加了以下条件判断和调用：

```
if (
    self.fd_config.speculative_config.method == SpecMethod.MTP
    and hasattr(self.proposer.model, "empty_input_forward")
    and self.parallel_config.use_ep
):
    self._execute_empty_mtp_input(self.forward_meta)
```

关键改动点：

1. 条件判断：检查 speculative method 是否为 MTP、proposer.model 是否具有 empty_input_forward 方法、是否启用 EP。
2. 函数调用：满足条件时调用 _execute_empty_mtp_input，与 execute_model_normal 中的逻辑对齐。

评论区精华

review 中只有 fastdeploy-bot 的 AI 代码审查评论，指出 PR 描述模板未填写具体内容，并建议补充背景和修改说明：

> 在 overlap schedule 模式下，当 real_bsz == 0 或 model_output is None 时，没有调用 MTP 的空输入处理函数，导致 EP 模型出现空运行问题。与 execute_model_normal 中的逻辑不一致。

该评论未引发进一步讨论，yuanlehome 直接批准了 PR。

风险与影响

风险分析：

- 测试覆盖不足：根据 codecov 报告，新增的 6 行代码 patch coverage 为 0%，缺少单元测试验证。
- 条件依赖：逻辑依赖于多个运行时状态（speculative method、model 方法、EP 启用），若状态异常可能引发错误。

影响分析：

- 影响范围：仅影响使用 overlap schedule、MTP 推测解码和 EP 模型的场景。
- 影响程度：修复后空输入处理恢复正常，避免潜在的空运行问题，提升系统可靠性。对用户无直接感知影响。

关联脉络

从近期历史 PR 看，本次 PR 属于 BugFix 类别，与 PR#7221（修复 GPU 异步拷贝和 Flash Mask Attention bug）类似，都是针对特定场景的底层逻辑修复。未发现直接关联的 PR 或 Issue，但可视为对调度和引擎模块的持续优化的一部分。