PR #1722 完整报告

THUDM/slime

[docker] patches for glm4.6v, kimi k2.5 and dsa cp only

合并时间：2026-03-13 15:16

原文链接：http://prhub.com.cn/THUDM/slime/pull/1722

PR 分析报告

执行摘要

本次 PR 更新了 Docker 补丁以支持新模型 GLM4.6V 和 Kimi K2.5，并优化了分布式训练中的 CP 逻辑，提升系统兼容性和稳定性。

功能与动机

动机源于添加对新发布模型的支持并修复分布式训练配置问题。PR body 为空，但从标题推断，旨在确保系统能够正确处理这些模型的训练和推理，解决潜在的死锁和兼容性问题。

实现拆解

- docker/patch/latest/megatron.patch：修改 MultimodalRotaryEmbedding 类，添加 packed_seq 判断，优化 CP 切片逻辑以避免在 THD 格式下重复处理。
- docker/patch/latest/sglang.patch：调整 ModelConfig 中的错误处理，将 Transformers 版本不兼容错误从 ValueError 降级为 warning，允许兼容旧版本；在 SchedulerDisaggregationPrefillMixin 中引入超时处理和调整 gloo 分组（使用 full TP group 当 CP > 1）以防止死锁。
- docker/version.txt：更新版本号至 nightly-dev-20260313a，反映补丁更新。

评论区精华

本次 PR 没有 review 评论或讨论，因此无争议点或决策记录。

风险与影响

风险：修改 CP 逻辑可能引入回归错误，影响注意力计算正确性；错误处理降级可能导致在不兼容版本下运行而未及时报错；超时配置需要合理设置，否则可能导致请求过早失败。 影响：对用户而言，可直接使用新模型进行训练；系统分布式训练更稳定；团队需持续维护 Docker 补丁更新。

关联脉络

与近期 PR 1743（更新 sglang 补丁）、PR 1745（GLM4V 多模态支持改进）和 PR 1749（GLM 模型配置更新）相关，显示团队持续优化模型支持和 Docker 基础设施，形成对 GLM 系列模型和多模态训练的持续演进。