PR #21192 完整报告

sgl-project/sglang

Fix CP in-seq-split method for DeepSeek V32 and update related tests

合并时间：2026-03-24 03:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21192

执行摘要

本 PR 修复了 DeepSeek V3.2 模型在上下文并行模式下 in-seq-split 方法的计算错误，通过调整 attn_cp_size 计算并更新测试使用官方模型，提升推理准确性和测试环境一致性。

功能与动机

动机源于修复 CP in-seq-split 方法中的 attn_cp_size 计算错误。根据 PR body，目标是修复 in-seq-split for CP of V32，并移除实验性模型引用，使用官方 DeepSeek-V3.2 进行测试，以提高测试可靠性和维护性。

实现拆解

主要改动点：

1. 核心计算修复：在 python/sglang/srt/server_args.py 的 _handle_model_specific_adjustments 方法中，将 self.attn_cp_size = self.tp_size 改为 self.attn_cp_size = self.tp_size // self.dp_size，修复了当数据并行启用时的计算错误。
2. 测试路径更新：多个测试文件（如 test/manual/nightly/test_deepseek_v32_perf.py、test/registered/8-gpu-models/test_deepseek_v32_basic.py 等）将模型路径从 "deepseek-ai/DeepSeek-V3.2-Exp" 替换为 "deepseek-ai/DeepSeek-V3.2"，统一使用官方模型版本。
3. 测试重组：移除 test/registered/8-gpu-models/test_deepseek_v32_cp_single_node.py，并新增 test/registered/cp/test_deepseek_v32_cp_single_node.py，将 CP 测试迁移到 cp 文件夹并标记为 pr-test，优化测试组织。

评论区精华

review 中仅有一个评论来自 gemini-code-assist[bot]：

> "The model path now points to the non-experimental DeepSeek-V3.2 model. For clarity and consistency with other test files, consider renaming the variable DEEPSEEK_V32_EXP_MODEL_PATH to DEEPSEEK_V32_MODEL_PATH and updating its usages in this file." 讨论焦点是代码风格一致性，无争议点，状态为已解决，PR 已合并。

风险与影响

风险：attn_cp_size 计算错误可能导致上下文并行效率降低或推理错误；测试路径更新可能意外引入依赖问题，但 CI 测试覆盖确保正确性。 影响：对使用 DeepSeek V3.2 进行 CP 推理的用户，修复后准确性提升；系统层面仅影响特定模型模式；团队测试环境更标准化，减少维护实验模型的开销。

关联脉络

与历史 PR #21170 "Fix CP residual size mismatch crash when tp_size == attn_cp_size" 相关，两者都针对上下文并行（CP）的 bugfix，显示团队在持续优化并行计算逻辑。