

PR #35588 完整报告

sgl-project/sglang

[Bugfix] Fix full prefill CUDA graph padding and EAGLE capture

合并时间: 2026-08-31 12:30

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35588>

执行摘要

- 一句话:
- 推荐动作:

功能与动机

实现拆解

1. 核心源码路径 - `python/sglang/srt/model_executor/runner/prefill_cuda_graph_runner.py` (data-contract) : 源码主路径; 包含 控制流调整、配置键调整; +15/-6 - `python/sglang/srt/model_executor/cuda_graph_buffer_registry.py` (data-contract) : 源码主路径; 包含 控制流调整、配置键调整; +12/-7 - `python/sglang/srt/model_executor/model_runner_components/cuda_graph_setup.py` (data-contract) : 源码主路径; +4/-4; 以 data-contract 为主
2. 测试与验证 - `test/registered/cuda_graph/full_prefill/test_full_cuda_graph_prefill.py` (test-coverage) : 测试配套; 涉及符号 `_prefill_graph_count`, `TestFullCudaGraphPrefillWithEagle3`, `test_eagle_target_prefill_replays_full_cuda_graph`; 包含 测试覆盖调整、导入关系调整、配置键调整; +67/-5; 关键符号: `_prefill_graph_count`, `TestFullCudaGraphPrefillWithEagle3`, `test_eagle_target_prefill_replays_full_cuda_graph` - `test/registered/unit/model_executor/test_prefill_cuda_graph_runner.py` (test-coverage) : 测试配套; 涉及符号 `test_eagle_target_full_reaches_graph_construction`; 包含 测试覆盖调整、配置键调整、符号定义调整; +39/-12; 关键符号: `test_eagle_target_full_reaches_graph_construction` - `test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_hybrid_bit_exact.py` (test-coverage) : 测试配套; 涉及符号 `_base_args`, `_prefill_graph_count`; 包含 测试覆盖调整、导入关系调整、配置键调整; +30/-3; 关键符号: `_base_args`, `_prefill_graph_count`

关键文件:

- `test/registered/cuda_graph/full_prefill/test_full_cuda_graph_prefill.py` (模块 `full/cuda/graph/prefill`; 类别 `test`; 类型 `test-coverage`; 符号 `_prefill_graph_count`, `TestFullCudaGraphPrefillWithEagle3`, `test_eagle_target_prefill_replays_full_cuda_graph`) : 测试配套; 涉及符号 `_prefill_graph_count`, `TestFullCudaGraphPrefillWithEagle3`,

test_eagle_target_prefill_replays_full_cuda_graph; 包含 测试覆盖调整、导入关系调整、配置键调整; +67/-5

- python/sglang/srt/model_executor/runner/prefill_cuda_graph_runner.py (模块 sglang/srt; 类别 source; 类型 data-contract) : 源码主路径; 包含 控制流调整、配置键调整; +15/-6
- python/sglang/srt/model_executor/cuda_graph_buffer_registry.py (模块 sglang/srt; 类别 source; 类型 data-contract) : 源码主路径; 包含 控制流调整、配置键调整; +12/-7
- test/registered/unit/model_executor/test_prefill_cuda_graph_runner.py (模块 prefill/cuda/graph/runner; 类别 test; 类型 test-coverage; 符号 test_eagle_target_full_reaches_graph_construction) : 测试配套; 涉及符号 test_eagle_target_full_reaches_graph_construction; 包含 测试覆盖调整、配置键调整、符号定义调整; +39/-12
- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_hybrid_bitexact.py (模块 unified/radix/cache/kl/hybrid/bitexact; 类别 test; 类型 test-coverage; 符号 _base_args, _prefill_graph_count) : 测试配套; 涉及符号 _base_args, _prefill_graph_count; 包含 测试覆盖调整、导入关系调整、配置键调整; +30/-3

关键符号: _prefill_graph_count, TestFullCudaGraphPrefillWithEagle3, test_eagle_target_prefill_replays_full_cuda_graph, test_eagle_target_full_reaches_graph_construction, _base_args, test_non_gathered_keeps_plain_fb_copy, test_non_gathered_uses_raw_token_count

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险:
- 影响:
 - 风险标记: 暂无

关联脉络

- 暂无明显关联 PR