

PR #36248 完整报告

sgl-project/sglang

[PP] Support prefill CUDA graph proxy tensors

合并时间: 2026-08-31 14:11

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36248>

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; 涉及符号 `_build_layer_model_forward_kwargs`, `_static_pp_proxy_tensors`, `_capture_pp_proxy_tensors`; 包含 控制流调整、配置键调整、符号定义调整; +40/-10; 关键符号: `_build_layer_model_forward_kwargs`, `_static_pp_proxy_tensors`, `_capture_pp_proxy_tensors` - `python/sglang/srt/model_executor/runner_utils/buffers.py` (data-contract) : 源码主路径; 涉及符号 `_allocate_pp_proxy_tensors`; 包含 控制流调整、配置键调整、符号定义调整; +59/-46; 关键符号: `_allocate_pp_proxy_tensors` - `python/sglang/srt/arg_groups/memory_hook.py` (dependency-wiring) : 源码主路径; 包含 导入关系调整、控制流调整、配置键调整; +15/-1
2. 测试与验证 - `test/registered/unit/model_executor/test_prefill_cuda_graph_runner_helpers.py` (test-coverage) : 测试配套; 涉及符号 `_make_pp_buffers_and_registry`, `test_pp_proxy_stable_buffers_accept_full_and_hidden_only_contracts`, `test_layer_model_kwargs_bind_optional_inputs_by_signature`, `proxy_before_embeds`; 包含 测试覆盖调整、导入关系调整、配置键调整; +127/-0; 关键符号: `_make_pp_buffers_and_registry`, `test_pp_proxy_stable_buffers_accept_full_and_hidden_only_contracts`, `test_layer_model_kwargs_bind_optional_inputs_by_signature`, `proxy_before_embeds` - `test/registered/unit/server_args/test_server_args.py` (test-coverage) : 测试配套; 涉及符号 `TestPipelineParallelPrefillCudaGraphPolicy`, `test_pp_prefill_graph_is_opt_in`, `test_pp_prefill_capture_limit_policy`; 包含 测试覆盖调整、导入关系调整、控制流调整; +54/-0; 关键符号: `TestPipelineParallelPrefillCudaGraphPolicy`, `test_pp_prefill_graph_is_opt_in`, `test_pp_prefill_capture_limit_policy` - `test/registered/unit/model_executor/model_runner_components/test_layer_setup.py` (test-coverage) : 测试配套; 涉及符号 `test_pipeline_placeholders_preserve_global_layer_ids`; 包含 测试覆盖调整、配置键调整、符号定义调整; +14/-0; 关键符号:

test_pipeline_placeholders_preserve_global_layer_ids

关键文件:

- python/sglang/srt/model_executor/runner/prefill_cuda_graph_runner.py (模块 sglang/srt; 类别 source; 类型 data-contract; 符号 _build_layer_model_forward_kwargs, _static_pp_proxy_tensors, _capture_pp_proxy_tensors): 源码主路径; 涉及符号 _build_layer_model_forward_kwargs, _static_pp_proxy_tensors, _capture_pp_proxy_tensors; 包含 控制流调整、配置键调整、符号定义调整; +40/-10
- python/sglang/srt/model_executor/runner_utils/buffers.py (模块 sglang/srt; 类别 source; 类型 data-contract; 符号 _allocate_pp_proxy_tensors): 源码主路径; 涉及符号 _allocate_pp_proxy_tensors; 包含 控制流调整、配置键调整、符号定义调整; +59/-46
- test/registered/unit/model_executor/test_prefill_cuda_graph_runner_helpers.py (模块 prefill/cuda/graph/runner/helpers; 类别 test; 类型 test-coverage; 符号 _make_pp_buffers_and_registry, test_pp_proxy_stable_buffers_accept_full_and_hidden_only_contracts, test_layer_model_kwargs_bind_optional_inputs_by_signature, proxy_before_embeds): 测试配套; 涉及符号 _make_pp_buffers_and_registry, test_pp_proxy_stable_buffers_accept_full_and_hidden_only_contracts, test_layer_model_kwargs_bind_optional_inputs_by_signature, proxy_before_embeds; 包含 测试覆盖调整、导入关系调整、配置键调整; +127/-0
- test/registered/unit/server_args/test_server_args.py (模块 server/args; 类别 test; 类型 test-coverage; 符号 TestPipelineParallelPrefillCudaGraphPolicy, test_pp_prefill_graph_is_opt_in, test_pp_prefill_capture_limit_policy): 测试配套; 涉及符号 TestPipelineParallelPrefillCudaGraphPolicy, test_pp_prefill_graph_is_opt_in, test_pp_prefill_capture_limit_policy; 包含 测试覆盖调整、导入关系调整、控制流调整; +54/-0
- python/sglang/srt/arg_groups/memory_hook.py (模块 sglang/srt; 类别 source; 类型 dependency-wiring): 源码主路径; 包含 导入关系调整、控制流调整、配置键调整; +15/-1

关键符号: _build_layer_model_forward_kwargs, _static_pp_proxy_tensors, _capture_pp_proxy_tensors, _allocate_pp_proxy_tensors, _make_pp_buffers_and_registry, test_pp_proxy_stable_buffers_accept_full_and_hidden_only_contracts, test_layer_model_kwargs_bind_optional_inputs_by_signature, proxy_before_embeds, test_finalize_pp_proxy_trims_padded_token_rows, test_bcg_eager_tail_uses_live_multimodal_embeddings

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