PR #24865 完整报告

sgl-project/sglang

speculative: drop dead params/returns/no-ops

合并时间：2026-05-10 06:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24865

执行摘要

- 一句话：清理推测解码模块死代码
- 推荐动作：推荐合入。此类死代码清理具有正向价值——降低认知负担、减少 Linter/Type Checker 误报、便于未来重构。可以快速 review 后合并。

功能与动机

干净代码是长期可维护性的基础。PR body 明确说明『Pure dead-code cleanup in speculative decoding workers and forward_batch_info: 4 atomic commits, no behavior change』，旨在通过消除死参数、未使用返回值和空操作来提升代码可读性并降低后续重构成本。

实现拆解

变更由 4 个原子提交组成，每个提交聚焦单一清理目标：

1. 移除冗余赋值(eagle_draft_extend_cuda_graph_runner.py)：删除了 spec_info.positions = None，该赋值已被前面逻辑覆盖，属于 no-op。
2. 删除 verify() 未使用的返回值(eagle_worker.py, multi_layer_eagle_worker.py, frozen_kv_mtp_worker.py)：验证方法 verify() 原本返回四元组 (logits_output, res, model_worker_batch, can_run_cuda_graph)，但 model_worker_batch 在所有调用点均被忽略（赋值给 _）。PR 将其改为返回三元组，并同步更新了所有调用处。
3. 简化 enable_num_token_non_padded 函数签名(forward_batch_info.py, cuda_graph_runner.py, cpu_graph_runner.py)：该函数原接受 server_args 参数但实际只使用了全局变量 get_moe_expert_parallel_world_size()，参数完全多余。PR 移除了参数并更新了所有调用处（包括两个 CUDA Graph runner 中的 3 处引用）。
4. 简化 check_forward_draft_extend_after_decode 方法签名(eagle_worker.py, multi_layer_eagle_worker.py)：该方法原接受 batch 和 verify_output 两个参数，但方法体内只用到了 verify_output 和 self.server_args，batch 参数未被使用。PR 移除了 batch 参数并更新调用处。

无测试或配置变更，因为作者声明『no behavior change』，属于纯重构。

关键文件：

- python/sglang/srt/speculative/eagle_worker.py（模块 推测解码；类别 source；类型 core-logic；符号 check_forward_draft_extend_after_decode, verify, forward_batch_generation）：核心推测解码 worker，清理了 verify() 返回值中的 model_worker_batch 和 check_forward_draft_extend_after_decode 方法中多余的 batch 参数
- python/sglang/srt/speculative/multi_layer_eagle_worker.py（模块 推测解码；类别 source；类型 core-logic；符号 check_forward_draft_extend_after_decode, verify, forward_batch_generation）：多层 Eagle worker，与 eagle_worker.py 相同的死代码清理
- python/sglang/srt/model_executor/forward_batch_info.py（模块 前向批处理；类别 source；类型 data-contract；符号 enable_num_token_non_padded）：定义 enable_num_token_non_padded 函数并本 PR 移除了其 server_args 参数
- python/sglang/srt/model_executor/cuda_graph_runner.py（模块 CUDA Graph；类别 source；类型 data-contract）：两处调用 enable_num_token_non_padded 跟随签名变更
- python/sglang/srt/model_executor/cpu_graph_runner.py（模块 CUDA Graph；类别 source；类型 data-contract）：一处调用 enable_num_token_non_padded 跟随签名变更
- python/sglang/srt/speculative/frozen_kv_mtp_worker.py（模块 推测解码；类别 source；类型 core-logic）：移除 verify() 返回值中未使用的 model_worker_batch
- python/sglang/srt/speculative/eagle_draft_extend_cuda_graph_runner.py（模块 推测解码；类别 source；类型 core-logic）：删除冗余赋值 spec_info.positions = None

关键符号：enable_num_token_non_padded, check_forward_draft_extend_after_decode, verify, forward_batch_generation

关键源码片段

python/sglang/srt/speculative/eagle_worker.py

核心推测解码 worker，清理了 verify() 返回值中的 model_worker_batch 和 check_forward_draft_extend_after_decode 方法中多余的 batch 参数

```
# eagle_worker.py 变更摘要（关键符号：verify, check_forward_draft_extend_after_decode）

# 1. verify() 返回值从四元组简化为三元组，移除了未使用的 model_worker_batch
# 旧：return logits_output, res, model_worker_batch, can_run_cuda_graph
# 新：return logits_output, res, can_run_cuda_graph

# 2. check_forward_draft_extend_after_decode 移除未使用的 batch 参数
# 旧：def check_forward_draft_extend_after_decode(self, batch: ScheduleBatch, verify_output: EagleVerifyOutput):
# 新：def check_forward_draft_extend_after_decode(self, verify_output: EagleVerifyOutput):

# 3. 调用 verify() 处同步解包
# 旧：logits_output, verify_output, model_worker_batch, can_run_cuda_graph = self.verify(batch, spec_info)
# 新：logits_output, verify_output, can_run_cuda_graph = self.verify(batch, spec_info)
```

python/sglang/srt/model_executor/forward_batch_info.py

定义 enable_num_token_non_padded 函数并本 PR 移除了其 server_args 参数

```
# forward_batch_info.py 中 enable_num_token_non_padded 函数签名变更
# 旧：def enable_num_token_non_padded(server_args):
# return get_moe_expert_parallel_world_size() > 1
# 新：def enable_num_token_non_padded():
# return get_moe_expert_parallel_world_size() > 1
# 注：server_args 参数实际未被使用，作为死参数移除。
```

评论区精华

该 PR 无实质性 review 讨论（0 条 review 评论），PR body 和提交信息已清晰说明每步变更的目的。仅有的 4 条 Issue 评论均为 CI 自动回复和 /rerun-test 命令。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。所有变更均是删除未使用的代码（参数、返回值、赋值），不影响运行时行为。关键风险点：
  - 若 model_worker_batch 将来被需要，需重新添加，但可通过 git revert 轻松恢复。
  - 若其他分支或未合入的 PR 引用了这些被删除的参数，可能产生合并冲突。
  - 影响：影响面窄但正面：仅影响推测解码模块（EagleWorker、MultiLayerEagleWorker、FrozenKvMtpWorker）及前向批处理信息模块（ForwardBatchInfo）。对用户无感知，对开发者则降低了阅读和维护成本。
  - 风险标记：暂无

关联脉络

- PR #24735 [Spec] Move accept_tokens off EagleDraftInput; pass via method arg: 同为 speculative decoding 模块的重构，涉及 EagleDraftInput 和 worker 方法签名变更，体现持续改善代码质量的趋势