

PR #30977 完整报告

sgl-project/sglang

[Spec] Rename `num\_tokens\_per\_bs` to `num\_tokens\_per\_req`

合并时间: 2026-07-14 02:47

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30977>

执行摘要

- 一句话: 机械重命名 `num_tokens_per_bs` 为 `num_tokens_per_req`
- 推荐动作: 建议快速合并。该 PR 是纯命名改进, 不引入行为变更, 能提升 speculative 解码路径的代码一致性, 值得合入。

功能与动机

PR 作者指出变量名 `num_tokens_per_bs` 具有误导性: 该值实际是每个请求槽位的 token 计数, 而非每 batch。EagleVerifyInput 已使用正确的 `num_tokens_per_req`, 因此统一重命名以消除歧义。

实现拆解

1. 核心接口重命名: 在 `spec_info.py` 和 `spec_registry.py` 中将 `get_num_tokens_per_bs_for_target_verify` 重命名为 `get_num_tokens_per_req_for_target_verify`, 并添加带 `DeprecationWarning` 的旧名别名。
2. 消费端适配: 将所有调用该接口的方法和变量重命名, 包括 `model_runner.decode_num_tokens_per_bs -> decode_num_tokens_per_req`, `_alloc_dummy_decode_buffers` 参数 `num_tokens_per_bs -> num_tokens_per_req`, 以及 `decode_cuda_graph_runner` 等中的同名属性。
3. 注释更新: 同步更新所有文档字符串和注释, 将引用改为新名。
4. 测试适配: 修改所有测试文件中的对应变量, 确保功能等价。
5. 额外清理: 初始提交还移除了 `SGLANG_ENABLE_SPEC_V2` 环境变量引用, 但后续分离到独立 PR, 最终只包含 `rename`。

关键文件:

- `python/sglang/srt/model_executor/model_runner.py` (模块 模型执行器; 类别 source; 类型 data-contract; 符号 `decode_num_tokens_per_bs`, `decode_num_tokens_per_req`): 核心入口文件, 暴露 `decode_num_tokens_per_req` 方法, 调度 speculative 和 normal 分支。
- `python/sglang/srt/speculative/spec_info.py` (模块 推测信息; 类别 source; 类型 core-logic; 符号 `get_num_tokens_per_bs_for_target_verify`, `get_num_tokens_per_req_for_target_verify`): 定义 `get_num_tokens_per_req_for_target_verify` 及其 deprecated 别名, 是 `rename` 的核心接

口层。

- python/sglang/srt/speculative/spec_registry.py (模块 算法注册; 类别 source; 类型 core-logic; 符号 get_num_tokens_per_bs_for_target_verify, get_num_tokens_per_req_for_target_verify) : CustomSpecAlgo 基类同样改名, 确保插件兼容。
- python/sglang/srt/model_executor/runner/base_runner.py (模块 基础执行器; 类别 source; 类型 data-contract; 符号 _alloc_dummy_decode_buffers) : _allocate_decode_buffers 和 _alloc_dummy_decode_buffers 参数重命名, 影响 dummy forward 和 warmup 路径。
- python/sglang/srt/model_executor/runner/decode_cuda_graph_runner.py (模块 解码图执行器; 类别 source; 类型 data-contract) : CUDA graph 解码 runner 中使用 num_tokens_per_req 决定 bucket 尺寸和 forward 逻辑。

关键符号: decode_num_tokens_per_req, get_num_tokens_per_req_for_target_verify, _alloc_dummy_decode_buffers

关键源码片段

python/sglang/srt/speculative/spec_info.py

定义 get_num_tokens_per_req_for_target_verify 及其 deprecated 别名, 是 rename 的核心接口层。

```
import warnings
```

```
class SpeculativeAlgorithm(Enum):
```

```
    # ... ( 其他代码 ) ...
```

```
    def get_num_tokens_per_req_for_target_verify(
        self, num_draft_tokens: int, is_draft_worker: bool
    ) -> int:
        # FIXME: Remove this after the forward mode refactor.
        if self.is_dspark() and is_draft_worker:
            return num_draft_tokens - 1
        return num_draft_tokens
```

```
    def get_num_tokens_per_bs_for_target_verify(
        self, num_draft_tokens: int, is_draft_worker: bool
    ) -> int:
        # Deprecated alias; remove together with the FIXME above.
        warnings.warn(
            "get_num_tokens_per_bs_for_target_verify is deprecated; use "
            "get_num_tokens_per_req_for_target_verify instead.",
            DeprecationWarning,
            stacklevel=2,
        )
        return self.get_num_tokens_per_req_for_target_verify(
            num_draft_tokens, is_draft_worker
```

)

评论区精华

该 PR 无 review 评论，为作者自合。PR 描述说明是纯机械改名，已验证 byte-identical，无需讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。作者声明已验证 head 与全局文本替换后结果 byte-identical。改动涉及 38 个文件，但均为全局替换，没有新增逻辑。保留了 deprecated 别名，任何遗漏的旧名调用仍可工作并触发警告。最大风险是完全未重命名的引用，但通过全局替换和 CI 应能捕获。
- 影响：对用户无影响，所有公共接口兼容。对开发者影响：新代码应使用 `num_tokens_per_req` 命名，旧名将在未来版本中移除。该 rename 提升了代码可读性，降低了因命名误解引入 bug 的风险。
- 风险标记：大量文件变更，依赖全局替换正确性，需注意后续移除 deprecated 别名

关联脉络

- PR #31008 [Spec] Deduplicate spec-v2 worker lifecycle boilerplate into BaseSpecWorker: 同为 speculative 模块重构，涉及 worker 基类抽象，可能需与本 PR rename 对齐。