# PR #27552 完整报告

sgl-project/sglang

[Spec] Rename token resolver to `_resolve_spec_v2_tokens`; remove dead V1 helpers

合并时间：2026-06-09 05:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27552

## 执行摘要

- 一句话：重命名函数并删除废弃的 Spec V1 辅助代码
- 推荐动作：该 PR 为常规清理维护，值得简要了解其删除逻辑，确认无意外引用即可快速合并。不包含复杂设计决策，无需深入阅读。

## 功能与动机

该 PR 是在 #25464 废弃 Spec V1 后的进一步清理。原函数名 _resolve_spec_overlap_tokens 暗示仅用于 overlap 场景，但实现在 V2 中已被用于非 overlap 路径，命名具有误导性。与此同时，V1 阶段遗留的辅助函数和 Triton kernel 已无引用，需要移除以避免混淆。详见 PR body。

## 实现拆解

1. 重命名函数：在 batch_result_processor.py 中将 _resolve_spec_overlap_tokens 重命名为 _resolve_spec_v2_tokens，并更新文档字符串以反映新用途。
2. 移除 V1 辅助函数：删除 eagle_utils.py 中的 apply_eagle_prefill_input_rotation 函数（24 行）和 spec_utils.py 中的 get_last_loc_large_page_size_large_top_k 函数（40 行），并清理相关导入。
3. 移除 Triton kernel：删除 cache_locs.py 中的 assign_draft_cache_locs kernel（约 110 行），该 kernel 已被 V2 的 assign_draft_cache_locs_page_size_1 和分页 torch.gather 路径取代。
4. 移除对应测试：删除整个 test/manual/spec/test_spec_utils.py 文件（348 行），该文件仅测试被删除的 kernel。
5. 微调注释：在 weight_updater.py 中更新一条注释，将 DFlashWorker 扩展为 DFlash / FrozenKVMTP workers。

关键文件：

- python/sglang/srt/managers/scheduler_components/batch_result_processor.py（模块 调度器；类别 source；类型 core-logic；符号 _resolve_spec_v2_tokens, _resolve_spec_overlap_tokens）：核心方法重命名，反映了 spec-v2 在 overlap 和 non-overlap 路径的统一处理
- python/sglang/srt/speculative/eagle_utils.py（模块 推测解码；类别 source；类型 core-logic；符号 apply_eagle_prefill_input_rotation）：删除了 V1 辅助函数

apply_eagle_prefill_input_rotation，该函数在 V2 中已不再使用

- python/sglang/srt/speculative/spec_utils.py（模块 推测解码；类别 source；类型 core-logic；符号 get_last_loc_large_page_size_large_top_k）：删除了 V1 辅助函数 get_last_loc_large_page_size_large_top_k 及关联导入
- python/sglang/srt/speculative/triton_ops/cache_locs.py（模块 推测核；类别 infra；类型 infrastructure；符号 assign_draft_cache_locs）：删除了 V1 Triton kernel assign_draft_cache_locs，该 kernel 已被 V2 版本替代
- test/manual/spec/test_spec_utils.py（模块 测试；类别 test；类型 test-coverage；符号 TestSpecUtils, setUp, test_assign_draft_cache_locs_single_seq, test_assign_draft_cache_locs_multi_seq）：删除了仅用于测试已删除 kernel 的手动测试文件
- python/sglang/srt/managers/scheduler_components/weight_updater.py（模块 调度器；类别 source；类型 other）：更新注释以正确反映 DFlash 和 FrozenKVMTP 工人都暴露 draft_model_runner

关键符号：_resolve_spec_v2_tokens, apply_eagle_prefill_input_rotation, get_last_loc_large_page_size_large_top_k, assign_draft_cache_locs

## 关键源码片段

### python/sglang/srt/speculative/eagle_utils.py

删除了 V1 辅助函数 apply_eagle_prefill_input_rotation，该函数在 V2 中已不再使用

```
# 删除 apply_eagle_prefill_input_rotation 后，剩余的核心函数如下：

def per_step_draft_out_cache_loc(
    out_cache_loc: torch.Tensor,
    batch_size: int,
    topk: int,
    num_steps: int,
) -> torch.Tensor:
    """Per-step slice of the multi-step EAGLE draft out_cache_loc buffer."""
    expected = batch_size * topk * num_steps
    assert out_cache_loc.shape[0] == expected
    return (
        out_cache_loc.view(batch_size, topk, num_steps)
        .permute(2, 0, 1)
        .reshape(num_steps, -1)
    )

def _eagle_prefill_tail_tokens(
    batch: ScheduleBatch, next_token_ids: torch.Tensor
) -> torch.Tensor:
    """Per-seq tail token for EAGLE prefill rotation; uses next prompt token for
    non-final chunks (chunked-prefill chain consistency, see PR #26329)."""
    tail_tokens = next_token_ids.to(batch.input_ids.dtype)
    next_prompt_token = batch.chunked_req_next_prompt_token
```

```
    if next_prompt_token is not None:
        for i, r in enumerate(batch.reqs):
            if r is batch.chunked_req:
                tail_tokens = tail_tokens.clone()
                tail_tokens[i] = next_prompt_token
                break
    return tail_tokens
```

## 评论区精华

该 PR 没有实质性的人工 review 讨论。仅有一个自动 bot 评论指出重命名和注释更新，无争议或权衡。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险极低。所有被删除的代码在 V2 中均无引用，并通过测试验证（见 CI 中重跑的 spec 测试均通过）。唯一的风险是如果未来有人尝试恢复 V1 可能需重新添加这些代码，但这属于设计决策，非本 PR 风险。
- 影响：对用户无功能影响。对开发者而言，减少了约 500 行死代码，提升了代码库整洁度。仅影响 speculative decoding 模块的内部实现，不影响其他模块。
- 风险标记：低风险，死代码删除

## 关联脉络

- PR #25464 [Spec] Deprecate Spec V1: 本 PR 是 #25464 废弃 Spec V1 后的后续清理，直接依赖其移除 V1 Worker 的决策
- PR #27599 [Spec] Naming cleanup: contiguous draft-loc kernel + accepted->accept: 同属于 speculative decoding 模块的命名清理系列 PR，互为补充