PR #27761 完整报告

sgl-project/sglang

[Spec] Remove dead `prepare_for_verify` / `prepare_extend_after_decode` + extend-decode kernel

合并时间：2026-06-12 11:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27761

执行摘要

- 一句话：清理推测解码 V1 残留死代码
- 推荐动作：建议快速审查当前 PR 确认无遗漏；对 speculative 架构演进感兴趣的开发者可结合 #27977 一起阅读，理解 V1→V2 的过渡细节。

功能与动机

PR body 指出：在 #27977 移除 V1 调度器分发和 EagleVerifyInput.verify / EagleVerifyOutput 后，这些辅助方法成为零引用死代码。保持代码库整洁，避免误导后续开发。

实现拆解

本 PR 分三步清理无引用代码：

1. 删除核心方法：在 python/sglang/srt/speculative/eagle_info.py 中移除 EagleVerifyInput.prepare_for_verify 和 EagleDraftExtendInput.prepare_extend_after_decode 两个方法，并删除相关的导入（ScheduleBatch、alloc_token_slots、assign_req_to_token_pool_func 等）。
2. 移除 Triton kernel：在 python/sglang/srt/speculative/triton_ops/cache_locs.py 中删除 create_extend_after_decode_spec_info Triton JIT kernel，该 kernel 仅服务于上一步的方法。
3. 清理再导出：在 python/sglang/srt/speculative/spec_utils.py 中移除对 create_extend_after_decode_spec_info 的再导出行。 无需额外测试，作者通过 ast + pyflakes 静态分析验证了零引用，原有 CI 测试（bookkeeping 所有权、frozen-KV MTP、EAGLE spec）均通过。

关键文件：

- python/sglang/srt/speculative/eagle_info.py（模块 推测解码；类别 source；类型 core-logic；符号 prepare_for_verify, prepare_extend_after_decode）：核心变更文件，删除两个里程碑级别的 V1 方法及大量相关导入。
- python/sglang/srt/speculative/spec_utils.py（模块 推测解码；类别 source；类型 dependency-wiring）：移除对已删除 Triton kernel 的再导出，保持导入关系清晰。

- python/sglang/srt/speculative/triton_ops/cache_locs.py（模块 推测解码；类别 infra；类型 infrastructure；符号 create_extend_after_decode_spec_info）：删除仅被上述方法调用的 Triton JIT kernel，减少维护负担。

关键符号：prepare_for_verify, prepare_extend_after_decode, create_extend_after_decode_spec_info

关键源码片段

python/sglang/srt/speculative/eagle_info.py

核心变更文件，删除两个里程碑级别的 V1 方法及大量相关导入。

```
@dataclass
class EagleVerifyInput(SpecInput, EagleVerifyInputV2Mixin):
    # 字段定义（与删除前一致）
    draft_token: torch.Tensor
    custom_mask: torch.Tensor
    positions: torch.Tensor
    retrieve_index: torch.Tensor
    retrieve_next_token: torch.Tensor
    retrieve_next_sibling: torch.Tensor
    retrieve_cum_len: torch.Tensor
    spec_steps: int
    topk: int
    draft_token_num: int
    capture_hidden_mode: CaptureHiddenMode
    seq_lens_sum: int
    seq_lens_cpu: torch.Tensor

    def __post_init__(self):
        super().__init__(SpecInputType.EAGLE_VERIFY)
        if self.num_tokens_per_req < 0:
            self.num_tokens_per_req = self.draft_token_num

    def get_spec_adjust_token_coefficient(self) -> Tuple[int, int]:
        return self.draft_token_num, self.draft_token_num

    @classmethod
    def create_idle_input(cls, topk, spec_steps, num_verify_tokens):
        '\'\'\'创建空闲批次桩输入（未修改）。\'\'\''
        return cls(
            draft_token=torch.empty((0,), dtype=torch.long, device='cuda'),
            custom_mask=torch.full((0,), True, dtype=torch.bool, device='cuda'),
            positions=torch.empty((0,), dtype=torch.int64, device='cuda'),
            retrieve_index=torch.full((0, num_verify_tokens), -1, dtype=torch.long, device='cuda'),
            retrieve_next_token=torch.full((0, num_verify_tokens), -1, dtype=torch.long, device=
            'cuda'),
            retrieve_next_sibling=torch.full((0, num_verify_tokens), -1, dtype=torch.long, device=
            'cuda'),
```

```
            retrieve_cum_len=None,
            topk=topk,
            draft_token_num=num_verify_tokens,
            spec_steps=spec_steps,
            capture_hidden_mode=CaptureHiddenMode.FULL,
            seq_lens_sum=0,
            seq_lens_cpu=torch.empty((0,), dtype=torch.int64),
        )

    # ========= 以下方法已被移除 =========
    # prepare_for_verify(self, batch, page_size):
    # V1 同步验证缓存分配方法，不再被任何调用者引用
    #  （所有推测算法已走 V2 worker）
```

评论区精华

PR 未引发实质性技术讨论。作者主动 /rerun-test 触发指定测试并通过，合并前无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。所有删除的方法及内核均经静态分析确认无任何调用点（ast + pyflakes）。保留的 EagleDraftExtendInput 字段仍被 FrozenKVMTPDraftExtendInput 读取，不受影响。CI 覆盖了关键 speculative 测试场景。
- 影响：
  - 用户：无功能变化，推理结果不受影响。
  - 系统：减少约 150 行代码和 1 个 Triton kernel，编译时间和二进制体积微降。
  - 团队：清理后 speculative 模块更聚焦于 V2 路径，降低代码理解成本。
  - 风险标记：低风险，静态验证，无功能影响

关联脉络

- PR #27977 [Spec] Remove the dead spec V1 scheduler paths: 本 PR 清理了该 PR 移除 V1 调度器后残留的辅助方法