PR #24735 完整报告

sgl-project/sglang

[Spec] Move `accept_tokens` off `EagleDraftInput`; pass via method arg

合并时间：2026-05-09 14:24

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24735

执行摘要

- 一句话：将 accept_tokens 从 EagleDraftInput 移除，改为方法参数传递
- 推荐动作：值得精读，尤其是理解 speculative decode 数据流如何逐步解耦。PR 设计上分离 draft input 与 verify output 的职责，是很好的架构演进方向。

功能与动机

在 #24724 中拆分了 verified_id 后，accept_tokens 仍然挂在 EagleDraftInput 上，导致 draft input 承载了 verify output 的职责。本 PR 将其彻底迁移到 EagleVerifyOutput 中，通过方法参数传递，使接口职责更单一。

实现拆解

1. 在 EagleVerifyOutput 中新增 create_idle 静态方法，统一 idle 场景下的输出创建，避免直接构造时传递大量空张量。
2. 从 EagleDraftInput 中移除 accept_tokens、seq_lens_for_draft_extend、seq_lens_for_draft_extend_cpu、req_pool_indices_for_draft_extend 等字段，转移到 EagleVerifyOutput 中。
3. 修改 forward_draft_extend_after_decode 和 check_forward_draft_extend_after_decode 方法，使其直接接收 EagleVerifyOutput 参数，从 verify_output 中读取 unfinished_accept_tokens 等信息，不再依赖 batch.spec_info.accept_tokens。
4. 在 verify 方法的返回处，将 draft_input 重命名为 next_draft_input，并调整 EagleVerifyOutput 的构造方式，同时将用于 extend 的序列信息直接放入 verify_output。
5. 同步更新 frozen_kv_mtp、dflash、ngram 等 info 类，移除遗留的 self.last_loc 赋值，并适配新的接口。

关键文件：

- python/sglang/srt/speculative/eagle_info.py（模块 推测解码；类别 source；类型 data-contract；符号 create_idle）：核心数据结构变更，移除了 EagleDraftInput.accept_tokens 等字段，新增 EagleVerifyOutput.create_idle 工厂方法，修改 verify 方法返回结构。
- 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,

forward_draft_extend_after_decode）：修改了 forward_draft_extend_after_decode 和 check_forward_draft_extend_after_decode 的签名与逻辑，从接收 batch.spec_info.accept_tokens 改为接收 verify_output.unfinished_accept_tokens。

- 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, forward_draft_extend_after_decode）：与 eagle_worker 同步修改，保持一致。
- python/sglang/srt/speculative/frozen_kv_mtp_worker.py（模块 推测解码；类别 source；类型 core-logic；符号 forward_draft_extend_after_decode）：适配新的接口，forward_draft_extend_after_decode 改为接收 verify_output，并从中获取 extend 所需字段。
- python/sglang/srt/speculative/frozen_kv_mtp_info.py（模块 推测解码；类别 source；类型 data-contract）：适配 EagleVerifyOutput.draft_input 重命名为 next_draft_input。
- python/sglang/srt/speculative/dflash_info.py（模块 推测解码；类别 source；类型 cleanup）：移除遗留的 self.last_loc 赋值。
- python/sglang/srt/speculative/ngram_info.py（模块 推测解码；类别 source；类型 cleanup）：移除遗留的 self.last_loc 赋值。

关键符号：EagleVerifyOutput.create_idle, EagleVerifyOutput.verify, forward_draft_extend_after_decode, check_forward_draft_extend_after_decode

关键源码片段

python/sglang/srt/speculative/eagle_info.py

核心数据结构变更，移除了 EagleDraftInput.accept_tokens 等字段，新增 EagleVerifyOutput.create_idle 工厂方法，修改 verify 方法返回结构。

```
class EagleInfo:
    def verify(self, batch: ScheduleBatch, logits_output: ...) -> EagleVerifyOutput:
        if batch.forward_mode.is_idle():
            # 空闲模式：创建空闲的 draft input，然后通过工厂方法构建 verify output
            next_draft_input = EagleDraftInput.create_idle_input(
                device=batch.device,
                hidden_size=batch.model_config.spec_hidden_size,
                dtype=batch.model_config.dtype,
                topk=self.topk,
                capture_hidden_mode=CaptureHiddenMode.LAST,
            )
            return EagleVerifyOutput.create_idle(
                next_draft_input=next_draft_input,
                logits_output=logits_output,
                device=batch.device,
                spec_steps=self.spec_steps,
            )
        # 非空闲分支省略 ...
```

评论区精华

该 PR 没有实质性的 review 讨论，仅包含 CI 测试触发指令和 bot 回复。

- 暂无高价值评论线程

风险与影响

- 风险：核心数据结构 EagleDraftInput 和 EagleVerifyOutput 的字段被重命名和移除，所有调用点（EagleWorker、MultiLayerEagleWorker、FrozenKVMTPWorker）已同步修改，但若存在未覆盖的第三方自定义 speculative 算法或分支，可能导致运行时属性缺失错误。另外，unfinished_accept_tokens 与原有 accept_tokens 的语义区分（子集 vs 全集）需要严格保证一致性，否则会影响 extend 前的数据准备。
- 影响：影响范围限于 speculative decoding 模块的 worker 和 info 类，用户无感知。内部数据流更清晰，为后续支持多 draft 或更复杂的 verify 逻辑打下基础。建议合并后运行 commited 测试套件（如 test_eagle_dp_attention.py、test_deepseek_v3_mtp.py）确保回归覆盖。
- 风险标记：核心数据结构变更，缺少直接测试覆盖，多 worker 同步风险

关联脉络

- PR #24724 [Spec] Disambiguate verified_id into bonus_token(s) / accept_tokens: 本 PR 是 #24724 的后续，进一步清理 accept_tokens 的使用和数据流。