PR #28093 完整报告

sgl-project/sglang

[Spec] Move draft-extend prep to `EagleDraftWorkerBase`; unify `prepare_for_*` names

合并时间：2026-06-13 07:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28093

执行摘要

- 一句话：将 draft-extend prep 移到 Worker 基类并统一方法命名
- 推荐动作：值得精读。该 PR 展示了如何通过重构消除架构异味（数据类不该持有方法）、统一命名规范，并体现了良好的小型重构实践。对于关注 speculative decoding 模块的开发者，这是熟悉模块结构的好机会。

功能与动机

PR 说明指出："Building the draft-extend forward is worker behavior, not data carried by the spec_info。" 因此将 prepare_for_extend_to_fill_draft_kvcache 从混入数据类移至 Worker 基类，更忠实于职责分离原则。

实现拆解

1. 移动 prepare_for_draft_extend 到基类：将 EagleDraftExtendInputV2Mixin.prepare_for_extend_to_fill_draft_kvcache 方法（byte-identical）复制到 EagleDraftWorkerBase 作为新方法 prepare_for_draft_extend，并将原方法中的 self 改为 draft_extend_input 参数；删除 EagleDraftExtendInputV2Mixin 类，EagleDraftExtendInput 不再继承它。涉及文件：base_spec_worker.py（新增方法）、eagle_info_v2.py（移除类和方法）、eagle_info.py（调整继承）。
2. 重命名基类：将 BaseDraftWorker 重命名为 EagleDraftWorkerBase，以准确反映其职责（仅 EAGLE 风格 draft worker 使用；NGRAM 没有 draft worker）。更新所有文件中的导入和类继承声明，包括 eagle_worker_v2.py、multi_layer_eagle_worker_v2.py、frozen_kv_mtp_worker_v2.py、ngram_worker.py、spec_utils.py、kv_canary/plan_input.py、dflash_worker_v2.py 等。
3. 统一 prep 方法命名：在 eagle_info_v2.py 中将 prepare_for_v2_draft 重命名为 prepare_for_draft，在 dflash_info.py 中将 prepare_for_v2_verify 重命名为 prepare_for_verify；在所有调用处（eagle_worker_v2.py、multi_layer_eagle_worker_v2.py、dflash_worker_v2.py 等）同步更新方法名和注释。
4. 更新 worker 调用：在 eagle_worker_v2.py 和 multi_layer_eagle_worker_v2.py 的 _draft_extend_for_decode 方法中，将原先 draft_extend_input.prepare_for_extend_to_fill_draft_kvcache(...) 改为 self.prepare_for_draft_extend(draft_extend_input, ...)；verify 方法中的调用也相应重命名。所有改动保持字节级别语义一致。

5. 测试适配：更新 test_decode_bookkeeping_ownership.py 中的类型守卫引用，将 BaseDraftWorker 替换为 EagleDraftWorkerBase，确保类型检查通过。

关键文件：

- python/sglang/srt/speculative/base_spec_worker.py（模块 spec 基类；类别 source；类型 core-logic；符号 BaseDraftWorker, EagleDraftWorkerBase, prepare_for_draft_extend, draft_worker）：核心变更文件：新增 prepare_for_draft_extend 方法并重命名类为 EagleDraftWorkerBase，承担 draft-extend forward batch 构建职责。
- python/sglang/srt/speculative/eagle_info_v2.py（模块 spec 信息；类别 source；类型 core-logic；符号 prepare_for_v2_draft, prepare_for_draft, EagleDraftExtendInputV2Mixin, prepare_for_extend_to_fill_draft_kvcache）：移除 EagleDraftExtendInputV2Mixin 及其方法；重命名 prepare_for_v2_draft 和 prepare_for_v2_verify，清理导入。
- python/sglang/srt/speculative/eagle_worker_v2.py（模块 eagle worker；类别 source；类型 core-logic；符号 EagleDraftWorker）：更新导入和类继承；在 _draft_extend_for_decode 和 verify 中改用基类的 prepare_for_draft_extend 方法及新的方法名。
- python/sglang/srt/speculative/multi_layer_eagle_worker_v2.py（模块 多层 eagle；类别 source；类型 core-logic；符号 MultiLayerEagleDraftWorker）：与 eagle_worker_v2.py 类似，更新导入、类继承和调用点。

关键符号：prepare_for_draft_extend, prepare_for_draft, prepare_for_verify

关键源码片段

python/sglang/srt/speculative/base_spec_worker.py

核心变更文件：新增 prepare_for_draft_extend 方法并重命名类为 EagleDraftWorkerBase，承担 draft-extend forward batch 构建职责。

```
# 基类从 BaseDraftWorker 重命名为 EagleDraftWorkerBase，明确 EAGLE 风格
class EagleDraftWorkerBase(ABC):
    # draft() 和 draft_extend() 保持抽象

    def prepare_for_draft_extend(
        self,
        draft_extend_input: EagleDraftExtendInput,
        batch: ScheduleBatch,
        predict: torch.Tensor,
        num_draft_tokens: int,
        draft_model_runner: Any,
        cuda_graph_runner: Any,
    ):
        # 原属于 EagleDraftExtendInputV2Mixin，因“构建 forward batch
        # 属于 worker 责任”而被迁移至此
        gpu_only = batch.seq_lens_cpu is None
```

```
        batch.spec_info = draft_extend_input
        batch.input_ids = predict
        maybe_detect_oob(...) # 输入校验
        if gpu_only:
            batch.prefix_lens = batch.seq_lens.to(torch.int32)
            batch.extend_lens = torch.full(
                (bs,), num_draft_tokens, dtype=torch.int32, device=batch.seq_lens.device
            )
        else:
            batch.prefix_lens = batch.seq_lens_cpu.tolist()
            batch.extend_lens = [num_draft_tokens] * bs
        forward_batch = ForwardBatch.init_new(batch, draft_model_runner)
        # 在 forward_batch 上增加 seq_lens 以反映 draft extend 的写入
        forward_batch.seq_lens = forward_batch.seq_lens + num_draft_tokens
        can_cuda_graph = cuda_graph_runner and cuda_graph_runner.can_run(forward_batch)
        if not batch.forward_mode.is_idle() and not can_cuda_graph:
            draft_model_runner.attn_backend.init_forward_metadata(forward_batch)
            if not is_npu() or can_cuda_graph:
                forward_batch.mark_forward_metadata_ready()
        return forward_batch
```

python/sglang/srt/speculative/eagle_info_v2.py

移除 EagleDraftExtendInputV2Mixin 及其方法；重命名 prepare_for_v2_draft 和 prepare_for_v2_verify，清理导入。

```
# eagle_info_v2.py — 重构后：不再引入 EagleDraftExtendInput 和 Any
from __future__ import annotations
from dataclasses import dataclass
from typing import TYPE_CHECKING # 移除了 Any

import torch
import torch.nn.functional as F
# ... 导入省略 ...

# EagleDraftExtendInputV2Mixin 类已被彻底移除
# 原 prepare_for_extend_to_fill_draft_kvcache 方法移至 base_spec_worker.py

# 方法名统一：prepare_for_v2_draft -> prepare_for_draft
def prepare_for_draft(
    self: EagleDraftInput,
    req_to_token_pool: ReqToTokenPool,
    batch: ScheduleBatch,
    cuda_graph_runner: EAGLEDraftCudaGraphRunner,
    draft_model_runner: ModelRunner,
    topk: int,
    num_steps: int,
):
    # 方法体未变，仅重命名
    if not batch.forward_mode.is_idle():
```

```
bs = len(batch.seq_lens)
page_size = batch.token_to_kv_pool_allocator.page_size
# ... 分配 cache 位置等
```

评论区精华

无实质性讨论；仅包含自动化 bot 的 CI 触发和状态更新（/rerun-test、/tag-and-rerun-ci）。

- 暂无高价值评论线程

风险与影响

- 风险：由于是纯重构（代码移动 + 重命名），运行时行为保持不变，回归风险较低。但涉及 speculative decoding 核心路径，若导入或调用点遗漏更新可能导致运行时异常。CI 中已运行 spec 相关测试（test_spec_eagle.py、test_frozen_kv_mtp.py、test_dflash.py、test_spec_ngram.py、test_spec_standalone.py）并全部通过，风险可控。
- 影响：对用户完全透明，无功能变化。系统代码结构更加清晰：draft-extend 的 forward batch 构建职责由 Worker 基类承担，数据类回归纯数据角色。对团队而言，新开发者更容易理解架构分层；后续若要添加新的 draft worker 变体，只需继承 EagleDraftWorkerBase 并实现抽象方法。
- 风险标记：核心路径变更，跨文件命名调整，回归风险低

关联脉络

- PR #28081 [refactor] Fold FrozenKVMTPCudaGraphRunner onto the shared DecodeCudaGraphRunner base: 同为 speculative-decoding 模块的重构，体现了持续改进架构的演进趋势
- PR #28096 [Spec] Fix EagleDraftWorker draft-extend attn backend assignment: 修复同一 worker 的 bug，与本 PR 的重命名间接相关