

# PR #25464 完整报告

sgl-project/sglang

[Spec] Deprecate Spec V1

合并时间: 2026-06-09 04:10

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/25464>

## 执行摘要

- 一句话: 废弃 Spec V1 单体 Worker, 统一使用 V2 Worker
- 推荐动作: 值得精读。该 PR 展示了在达到功能对等后安全删除大量旧代码的典型模式: 先逐步补齐 V2 功能, 确认 CI 覆盖后一次性清除 V1。scheduler.py 中 `_forward_isolation` 的泛化设计也值得借鉴——将 `overlap` 特定逻辑参数化, 保持同步 / 异步路径的统一。建议团队在类似技术债务清理时参考此流程。

## 功能与动机

V1 此前因 V2 功能不完整而保留。近期合入的 Tree drafting topk>1 支持 (#26866/#26997/#26972)、Mamba 模型支持 (#27463)、NaN 修复 (#27545) 以及防御性边界检查 (#27486) 补齐了 V2 的能力差距。在 V2 已达到功能对等且调度器可驱动同步路径的前提下, V1 实现已成为技术债务, 删除后可显著简化 speculative decoding 模块的维护与演进。

## 实现拆解

1. 删除 V1 Worker 文件 移除 `eagle_worker.py` (1356 行)、`multi_layer_eagle_worker.py` (821 行) 和 `standalone_worker.py` (121 行), 完全清除 `EAGLEWorker`、`MultiLayerEagleWorker`、`StandaloneWorker` 类及其所有方法。相关 CUDA Graph runner 和 attention backend 的 V1 专用初始化也被删除。
2. 统一 Worker 创建路径 在 `spec_info.py` 的 `SpeculativeAlgorithm.create_worker` 方法中, 移除 `enable_overlap` 条件分支: `EAGLE`、`EAGLE3`、`STANDALONE`、`MULTI_LAYER_EAGLE` 始终返回 V2 Worker 类 (`EAGLEWorkerV2`、`MultiLayerEagleWorkerV2`、`StandaloneWorkerV2`)。NGRAM 和 FROZEN\_KV\_MTP 保持原逻辑。
3. Scheduler 同步 V2 路径 在 `scheduler.py` 的 `run_batch` 方法中, 为 `is_spec_v2` 且非 `overlap` 模式增加同步分支: 调用 `resolve_forward_inputs` 准备输入, 在 `_forward_isolation` 上下文内执行 `forward_batch_generation`, 然后恢复 `spec_info/seq_lens` 状态并拷贝结果到 CPU。原有的 `_overlap_forward_isolation` 被泛化为 `_forward_isolation(batch, *, overlap)`, 非 `overlap` 路径跳过 2-iter tensor pinning (使用单流, 无需 `batch_record_buf`)。
4. 配套清理 - `kv_cache_builder.py`: 移除 `enable_overlap` 参数, V2 Worker 不再依赖 `overlap` 模式分配 draft KV pool。 - `speculative_hook.py`: 移除 V1 相关的配置钩子。 -

`debug_utils/pr_fix_toggle.py`: 删除只针对 V1 的补丁。 - `schedule_batch.py`: 移除 V1 特有的属性。

5. 测试适配 更新 `speculative_draft_runner.py` 中的测试辅助类, 使其始终使用 V2 Worker。注册的 spec 测试用例 (`test_spec_eagle.py` 等) 保持原样, CI 中增加了对同步 V2 路径的覆盖。

关键文件:

- `python/sglang/srt/speculative/eagle_worker.py` (模块 推测解码; 类别 source; 类型 deletion; 符号 `EAGLEWorker`, `init`, `init_attention_backend`, `init_cuda_graphs`): 被删除的核心 V1 Worker 文件, 包含 `EAGLEWorker` 类及其所有方法, 1356 行代码被移除。
- `python/sglang/srt/speculative/multi_layer_eagle_worker.py` (模块 推测解码; 类别 source; 类型 deletion; 符号 `MultiLayerEagleWorker`, `init`, `init_attention_backend`, `init_cuda_graphs`): 多 MTP 层 V1 Worker, 包含 `MultiLayerEagleWorker` 类及 `forward_batch_generation` 等关键方法, 821 行被移除。
- `python/sglang/srt/speculative/standalone_worker.py` (模块 推测解码; 类别 source; 类型 deletion; 符号 `StandaloneWorker`, `init`): `Standalone Worker` (继承自 `EAGLEWorker`) 被删除, 121 行代码移除。
- `python/sglang/srt/managers/scheduler.py` (模块 调度器; 类别 source; 类型 core-logic; 符号 `_overlap_forward_isolation`, `_forward_isolation`): 核心调度器新增同步 V2 路径, 并将 `_overlap_forward_isolation` 泛化为 `_forward_isolation`, 实现了 V2 Worker 在非 overlap 模式下的驱动。
- `python/sglang/srt/speculative/spec_info.py` (模块 推测解码; 类别 source; 类型 dependency-wiring): `SpeculativeAlgorithm.create_worker` 方法移除了 V1/V2 分支, 始终返回 V2 Worker, 简化了 Worker 创建逻辑。

关键符号: `_forward_isolation`, `create_worker`, `run_batch`

## 关键源码片段

### `python/sglang/srt/managers/scheduler.py`

核心调度器新增同步 V2 路径, 并将 `_overlap_forward_isolation` 泛化为 `_forward_isolation`, 实现了 V2 Worker 在非 overlap 模式下的驱动。

```
@contextmanager
```

```
def _forward_isolation(self, batch: ScheduleBatch, *, overlap: bool):
```

```
    """Make SB transactional across one forward (overlap and non-overlap).
```

```
    1. Snapshot SB fields so V2's mid-forward mutations can be undone.
```

```
    2. Substitute sampling_info with a forward-only copy.
```

```
    3. (overlap=True only) Pin (batch, snapshot) into batch_record_buf for 2 iters.
```

```
    """
```

```
    # 1. snapshot
```

```
    snapshot_v2_full = batch.is_spec_v2
```

```
    # 2. substitute sampling info
```

```
    sched_sampling_info = batch.sampling_info
```

```
    if sched_sampling_info is not None:
```

```

        batch.sampling_info = sched_sampling_info.copy_for_forward()
# 3. pin for 2-iter tensor lifetime (overlap path only)
if overlap:
    self.record_batch_in_overlap(batch)
try:
    yield
finally:
    # restore
    batch.is_spec_v2 = snapshot_v2_full
    if sched_sampling_info is not None:
        batch.sampling_info = sched_sampling_info

# In run_batch:
elif batch.is_spec_v2:
    # Non-overlap V2: drive V2 worker synchronously
    resolve_forward_inputs(batch, self.future_map)
    with self._forward_isolation(batch, overlap=False):
        batch_result = self.draft_worker.forward_batch_generation(batch)
    # re-apply spec_info and seq_lens
    ...
    batch.input_ids = None

```

### python/sglang/srt/speculative/spec\_info.py

`SpeculativeAlgorithm.create_worker` 方法移除了 V1/V2 分支，始终返回 V2 Worker，简化了 Worker 创建逻辑。

```

@staticmethod
def create_worker(
    algorithm: "SpeculativeAlgorithm",
    server_args: ServerArgs,
    enable_overlap: bool,
) -> type:
    # ... other algorithms ...
    if algorithm.is_eagle() and server_args.enable_multi_layer_eagle:
        # Always use V2 worker, no longer branches on enable_overlap
        from sglang.srt.speculative.multi_layer_eagle_worker_v2 import (
            MultiLayerEagleWorkerV2,
        )
        return MultiLayerEagleWorkerV2
    elif algorithm.is_eagle():
        from sglang.srt.speculative.eagle_worker_v2 import EAGLEWorkerV2
        return EAGLEWorkerV2
    elif algorithm.is_standalone():
        from sglang.srt.speculative.standalone_worker_v2 import StandaloneWorkerV2
        return StandaloneWorkerV2
    elif algorithm.is_ngram():
        # NGRAM worker unchanged
        ...

```

## 评论区精华

无实质性 review 讨论。PR 仅为作者自主推进的技术债务清理，评论中仅有 GitHub Actions 的自动回复和测试重跑指令。作者在 PR body 中详细说明了 V1 可删除的必要性，未受到质疑。

- 暂无高价值评论线程

## 风险与影响

- 风险：
  - 兼容性风险：V1 Worker 被完全删除，任何直接引用 EAGLEWorker、MultiLayerEagleWorker、StandaloneWorker 的外部代码将无法运行。官方 API 应已迁移到 V2，但需确认所有下游组件（包括第三方扩展）已完成适配。
  - NPU 支持：NPU 后端此前依赖 EAGLEWorker 和 MultiLayerEagleWorker 中的 NPU Graph Runner。PR 中修改了 eagle\_draft\_npu\_graph\_runner.py 和 eagle\_draft\_extend\_npu\_graph\_runner.py（仅 import 调整），但需要确认 V2 Worker 在 NPU 上功能正确。测试文件中包含了 NPU 标签，CI 已覆盖部分测试。
  - 回归风险：同步 V2 路径是新增代码路径，可能在边缘场景（如 adaptive speculative、EAGLE3 多 MTP 层）存在隐藏 bug。虽有关联修复合入，但缺乏独立的同步路径专项测试。
  - 性能影响：V2 Worker 设计用于 overlap 模式，同步路径下可能因缺少 overlap 而降低吞吐，但这是预期行为，不视为 bug。
- 影响：
  - 代码规模：净删除约 2.4k 行源码，scheduler.py 增加约 50 行，总体大幅降低维护成本。
  - 用户影响：所有使用 EAGLE/EAGLE3/STANDALONE/MULTI\_LAYER 的 speculative decoding 用户将自动使用 V2 Worker，无需改动配置。--disable-overlap-schedule 标志仍然有效，且现在能正确驱动 V2 Worker 同步运行。NGRAM 和 FROZEN\_KV\_MTP 不受影响。
  - 系统影响：Scheduler 顶部增加 is\_spec\_v2 判断的逻辑复杂度可忽略。V2 Worker 的同步路径与 overlap 路径共享核心实现，测试一致性好。
  - 团队影响：开发人员不再需要维护两套 Worker 实现，后续功能迭代只需聚焦 V2，降低认知负荷。
  - 风险标记：大量代码删除，核心路径变更，NPU 需验证，同步路径新增无专项测试

## 关联脉络

- PR #27545 Fix NaN in triton EAGLE spec-v2 draft-extend CUDA graph at topk>1 (wrong qo\_indptr stride): 该修复补齐了 V2 在 topk>1 时的功能对等，使 V1 可安全删除。
- PR #26866 Tree drafting topk>1 support on V2 (page\_size==1): V2 功能对等的关键补丁之一，PR body 中提及。
- PR #26997 Tree drafting topk>1 support on V2 (reland): V2 功能对等的关键补丁之一，PR body 中提及。

- PR #27463 Mamba/hybrid-linear model support on V2: V2 功能对等的关键补丁之一，PR body 中提及。
- PR #27486 Defensive guards for EAGLE draft KV indexing: V2 功能对等的关键补丁之一，PR body 中提及。