

PR #32110 完整报告

sgl-project/sglang

[Spec] Enable grammar overlap scheduling for STANDALONE speculative decoding

合并时间: 2026-07-23 16:47

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32110>

执行摘要

- 一句话: 为 STANDALONE 推理添加语法重叠调度
- 推荐动作: PR 改动简单, 值得快速合入。测试覆盖充分, 逻辑清晰。建议阅读 `spec_info.py` 中 `supports_grammar_overlap` 的设计方式, 了解 SGLang 如何通过枚举方法控制解码行为。

功能与动机

PR body 说明: "Enable the spec-v2 grammar barrier for STANDALONE (StandaloneWorkerV2 inherits the EAGLE V2 verify path where the barrier is threaded). Constrained-decoding coverage reuses the existing standalone triton server (overlap on), so grammar decode exercises the barrier path with no extra server launch." 即由于 StandaloneWorkerV2 已继承 EAGLE V2 的 verify 路径, 该路径已包含 grammar barrier, 因此只需在调度策略层面开启支持, 并补充测试覆盖。

实现拆解

1. 启用 grammar overlap 标志: 在 `python/sglang/srt/speculative/spec_info.py` 的 `SupportsGrammarOverlap` 方法中, 将 `return self.is_eagle()` 改为 `return self.is_eagle() or self.is_standalone()`, 并添加注释说明 STANDALONE 继承 EAGLE V2 worker verify 路径, barrier 已包含。
2. 补充约束解码测试: 在 `test/registered/spec/test_spec_standalone.py` 的 `TestStandaloneV2SpeculativeDecodingTriton` 类中, 混入 `RegexConstrainedMixin` 和 `JSONConstrainedMixin`, 并更新注释说明约束 Mixin 复用同一服务器且 overlap 开启, 从而验证 grammar barrier 路径。同时将注册的预估时间从 406 调整为 450, 反映新增测试用例的开销。

关键文件:

- `python/sglang/srt/speculative/spec_info.py` (模块 推理调度; 类别 source; 类型 core-logic; 符号 `SpeculativeAlgorithm`): 核心逻辑变更: 在 `SupportsGrammarOverlap` 中新增 `is_standalone()` 分支, 使 STANDALONE 算法能够参与 grammar barrier 调度。
- `test/registered/spec/test_spec_standalone.py` (模块 测试; 类别 test; 类型 test-coverage; 符号 `TestStandaloneV2SpeculativeDecodingTriton`): 测试配套: 在

Triton 后端的测试类中混入约束解码 Mixin，验证 grammar barrier 路径正确性。

关键符号: SpeculativeAlgorithm.supports_grammar_overlap

评论区精华

该 PR 没有 review 评论讨论。仅有一条 author 触发的 `/rerun-test` 命令和 bot 的结果报告，测试通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及两处，且核心逻辑只是新增一个条件分支。测试已覆盖 triton 后端的约束解码场景。但需注意：若未来 STANDALONE 的 verify 路径与 EAGLE 产生偏离，则此处的假设可能失效。
- 影响：影响范围：仅影响使用 STANDALONE 算法且开启 grammar overlap 的推理场景，默认行为不变。影响程度：较小，但提升了功能性——用户现在可以在 STANDALONE speculative decoding 中同时使用约束解码（如 JSON 或正则约束），而无需额外配置。
- 风险标记：假设依赖（STANDALONE 继承 EAGLE verify 路径）

关联脉络

- PR #32023 [Scheduler] Enable decode retraction ordering under speculative decoding: 同属 speculative decoding 调度优化，涉及 grammar overlap 上下文的 retraction 排序。