

PR #32211 完整报告

sgl-project/sglang

[CI] Fix Mamba ServerArgs namespace

合并时间: 2026-07-24 04:15

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32211>

执行摘要

- 一句话: 修复 Mamba ServerArgs 命名空间缺失
- 推荐动作: 建议快速合并。此 PR 属于常规维护性质, 修复了由 #31230 引入的 namespace 遗漏问题, 维持了 ServerArgs 命名空间覆盖检测的完整性。值得关注的是 NS() 声明机制在代码库中的一致使用规范。

功能与动机

PR #31230 新增了 `ServerArgs.mamba_max_states_per_path` 但没有添加 `NS("exec.mamba")` 元数据, 导致 `test_server_args_namespaces.py` 覆盖测试在 `main` 分支上失败, 因为新字段未被包含在生成的命名空间映射中。本 PR 旨在修复此问题, 恢复命名空间覆盖检测的正确性。

实现拆解

在 `python/sglang/srt/server_args.py` 文件中, `class ServerArgs` 下, `mamba_max_states_per_path` 字段定义的第三行 (帮助字符串之后) 添加了一行 `NS("exec.mamba")` 元数据。这一行将字段的命名空间归属声明为 `exec.mamba`, 与同一模块中的 `mamba_ssm_dtype`、`enable_mamba_cache_stochastic_rounding` 等字段保持一致。其他文件无任何修改。

关键文件:

- `python/sglang/srt/server_args.py` (模块 配置管理; 类别 `source`; 类型 `core-logic`; 符号 `mamba_max_states_per_path`): 唯一变更文件。在 `mamba_max_states_per_path` 字段定义中添加了 `NS("exec.mamba")` 元数据, 修复命名空间覆盖检测缺失导致的测试失败。

关键符号: 未识别

关键源码片段

`python/sglang/srt/server_args.py`

唯一变更文件。在 `mamba_max_states_per_path` 字段定义中添加了 `NS("exec.mamba")` 元数据, 修复命名空间覆盖检测缺失导致的测试失败。

```
# 在 class ServerArgs 的 Mamba 配置段中, mamba_max_states_per_path 字段定义如下:  
# 添加前的版本 (导致 bug):
```

```

mamba_max_states_per_path: A[
  int,
  "Maximum number of cached Mamba states retained per root-to-tail path "
  "(-1 means unlimited). When enabled, after each insert the shallowest eligible "
  "interior states beyond the cap are removed while their full KV remains. "
  "Tail, fork, and locked nodes are preserved. Must be -1 or a positive integer.",
  # 注意：此处缺少 NS("exec.mamba"), 导致命名空间覆盖检测失败
] = -1

# 添加后的版本（修复后）：
mamba_max_states_per_path: A[
  int,
  "Maximum number of cached Mamba states retained per root-to-tail path "
  "(-1 means unlimited). When enabled, after each insert the shallowest eligible "
  "interior states beyond the cap are removed while their full KV remains. "
  "Tail, fork, and locked nodes are preserved. Must be -1 or a positive integer.",
  NS("exec.mamba"), # <-- 添加此行，将该字段归入 exec.mamba 命名空间
] = -1

```

评论区精华

本次 PR 无实质性的 review 讨论。[ishandhanani](#) 询问是否可以合并，作者已通过 `/rerun-test` 命令重新运行了 namespace 测试并验证通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅添加一行命名空间声明，不影响运行时行为。回归风险几乎为零。
- 影响：影响范围极小，仅修复了 CPU 单元测试分区 `test_server_args_namespaces.py` 的失败。对于用户无感知，对系统无影响。
- 风险标记：暂无

关联脉络

- PR #31230 [Mamba] Add a per-path cap for cached states: 本 PR 修复了 #31230 引入的 bug：新增的 `ServerArgs.mamba_max_states_per_path` 缺少 `NS("exec.mamba")` 命名空间元数据。