PR #25238 完整报告

sgl-project/sglang

[CI] Bundle `check-changes` outputs + caller inputs into 2 JSON inputs

合并时间：2026-05-14 16:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25238

执行摘要

- 一句话：将 9 个独立输入打包为 2 个 JSON 输入，简化 CI 工作流调用
- 推荐动作：值得阅读。该 PR 展示了 GitHub Actions 工作流中通过 JSON 打包输入来简化多参数传递的设计模式，同时提供了验证等价性的方法。适合负责 CI 维护的工程师参考。

功能与动机

减少 CI 工作流中重复且冗余的参数传递，降低维护成本。PR body 指出原来每个 CUDA 阶段需要传递 9 个独立输入，通过打包为 JSON 可简化调用者存根，使工作流更简洁易维护。

实现拆解

1. 修改被调用工作流输入接口：在 _pr-test-stage.yml 中，将原有的 9 个输入（target_stage, test_parallel_dispatch, partitions, main_package, sgl_kernel, continue_on_error_flag, pr_head_sha, git_ref, skip_stage_health_check）合并为 3 个：check_changes（接收 check-changes.outputs 的 JSON）、caller_inputs（接收 inputs 的 JSON）、partitions（独立保留以避免矩阵表达式中双重 fromJson）。
2. 修改调用者工作流：在 pr-test.yml 中，所有 CUDA stage 的 with: 部分从逐个传递 9 个参数改为只传递 3 个，大幅减少重复代码。
3. 保留非 CUDA stage 的原始传递方式：arm/x64 wheel build 和 call-sgl-kernel-tests 等非 CUDA 阶段仍使用原始独立参数（通过最后两个 commit 恢复），避免影响非目标阶段。
4. 等价性验证：使用 Python 脚本将 GHA 上下文替换为固定 fixture 值，渲染两种形式的步骤列表并逐字节比较，通过 14/14 个 CUDA 阶段的等价性测试。

关键文件：

- .github/workflows/_pr-test-stage.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：定义可重用工作流的输入接口，是本次变更的核心：将 9 个独立输入合并为 3 个 JSON 包输入。
- .github/workflows/pr-test.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：调用者工作流，所有 CUDA stage 的调用参数被简化，体现变更对调用方的实际影响。

关键符号：未识别

关键源码片段

.github/workflows/_pr-test-stage.yml

定义可重用工作流的输入接口，是本次变更的核心：将 9 个独立输入合并为 3 个 JSON 包输入。

```
# _pr-test-stage.yml (after change)
# 输入接口：将原来独立传输的 check-changes 输出和调用者输入打包为 JSON。
# partitions 单独保留，避免矩阵表达式中的双重 fromJson。
on:
  workflow_call:
    inputs:
      self_name:
        description: 'Caller job key; used for partitions[suite] lookup and target_stage gating.'
        type: string
        required: true
      runner_config:
        description: 'Key in scripts/ci/runner_configs.yml (install script / artifact version / install
        timeout).'
        type: string
        required: true
      runs_on:
        description: 'GHA runner label. B200 stages pass needs.check-changes.outputs.b200_
        runner for dynamic selection.'
        type: string
        required: true
      # ---- 以下三个输入取代了原来的 9 个独立输入 ----
      check_changes:
        description: 'toJson(needs.check-changes.outputs). Read via fromJson(...).main_package /
        sgl_kernel / continue_on_error etc.'
        type: string
        required: true
      caller_inputs:
        description: 'toJson(inputs) from pr-test.yml. Read via fromJson(...).target_stage / pr_
        head_sha / git_ref / skip_stage_health_check / test_parallel_dispatch.'
        type: string
        required: true
      partitions:
        description: 'check-changes.outputs.partitions raw — kept separate to avoid double-
        fromJson in matrix expressions.'
        type: string
        required: true
```

评论区精华

无 review 评论，但 PR body 详细解释了将 partitions 保持独立的原因：由于 partitions 本身已经是 JSON 字符串，通过 toJson(needs.check-changes.outputs) 会变成转义字符串，导致矩阵表达式中需要双重 fromJson，而矩阵上下文无法使用步骤解析的值，因此将其保持为独立输入，简化矩阵表达式。

- partitions 输入为什么保持独立 (design): 接受该设计，partitions 保持独立输入。

风险与影响

- 风险：主要风险是 JSON 解析错误或参数映射遗漏导致 CI stage 配置错误。但作者通过等价验证脚本确认 14/14 CUDA 阶段行为一致，降低了回归风险。此外，非目标阶段（wheel build、sgl-kernel-tests）未受影响。
- 影响：对开发者：CI 工作流更简洁，后续添加新 stage 时只需传递 3 个参数。对用户：无直接影响。对系统：降低工作流维护成本，减少出错可能。
- 风险标记：JSON 解析风险，矩阵表达式中 fromJson 使用

关联脉络

- PR #25138 ci: extract cuda stage actions + runner_config mapping: PR #25138 将 CUDA stage 提取为可重用工作流，而 #25238 在此基础上进一步简化了输入传递方式，是前者的直接后续改进。