PR #24270 完整报告

sgl-project/sglang

[codex] Add official diffusion GT workflow mode

合并时间：2026-05-03 15:10

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24270

执行摘要

- 一句话：为扩散 CI 添加强制官方 GT 模式
- 推荐动作：如果负责扩散模型的 CI 维护，建议仔细阅读此 PR 以理解官方 GT 模式的工作原理；对其他开发者而言，了解即可，无需深入。

功能与动机

根据 PR 描述，需要引入一种可选的官方扩散 GT 模式，以运行官方可比案例进行一致性验证，作为原生 SGLang GT 生成的补充。这允许针对特定案例组进行定向 GT 再生，并能够将大型案例路由到 H200 以利用更多资源。

实现拆解

1. 新增输入参数：在 workflow 的 on.workflow_dispatch.inputs 中增加了三个输入项——run_official_cases（布尔，默认 false）、official_case_ids（空格分隔的案例 ID 字符串）、official_source_group（组筛选：all/diffusers/wan21/ltx23）。
2. 添加 compute-official-gt-matrix job：当 run_official_cases 为 true 时，运行一个 compute-official-gt-matrix job。该 job 使用内联 Python 脚本，根据 official_source_group 和 official_case_ids 计算要运行的案例矩阵，并将大型案例（如 Flux2 和 LTX 两阶段）标记为使用 H200 runner。矩阵作为 job 输出传递给后续的 main job。
3. 修改 main job：条件性地根据 run_official_cases 设置 PUBLISH_TARGET_DIR 默认值（官方输出到 diffusion-ci/consistency_gt/official_generated）。runner 选择逻辑被更新：如果矩阵中包含 H200 标签，则切换到 H200 runner。
4. 优化 artifact 上传：修改 artifact 上传步骤，只包含生成的图像和 manifest 文件，避免模型权重、输入资产等被错误上传作为 GT artifacts。
5. 更新并发组键：将新输入纳入并发组键计算，以避免相同 ref 下不同参数的工作流互相取消。

关键文件：

- .github/workflows/diffusion-ci-gt-gen.yml（模块 CI 工作流；类别 infra；类型 infrastructure；符号 link）：唯一的变更文件，新增官方扩散 GT 模式的核心实现，包括输入参数、矩阵计算 job、runner 路由和 artifact 优化。

关键符号：link

关键源码片段

.github/workflows/diffusion-ci-gt-gen.yml

唯一的变更文件，新增官方扩散 GT 模式的核心实现，包括输入参数、矩阵计算 job、runner 路由和 artifact 优化。

```
# 新增输入参数定义（工作流调度时可选）
run_official_cases:
  description: 'Run official comparable GT cases instead of native SGLang GT cases.'
  required: false
  default: false
  type: boolean
official_case_ids:
  description: 'Specific official case IDs to run (space-separated). Leave empty to run all official
  comparable cases.'
  required: false
  default: ''
  type: string
official_source_group:
  description: 'Official GT source group filter: all, diffusers, wan21, or ltx23. Used only when
  run_official_cases is true.'
  required: false
  default: ''
  type: string

# 并发组键包含新输入以避免冲突
concurrency:
  group: diffusion-ci-gt-gen-${{ github.ref }}-${{ inputs.output_name || inputs.case_ids || inputs.
  official_case_ids || inputs.official_source_group || inputs.run_official_cases || 'default' }}
  cancel-in-progress: true

# 根据 run_official_cases 决定默认发布目录
env:
  PUBLISH_TARGET_DIR: ${{ inputs.publish_target_dir || (inputs.run_official_cases && 'diffusion-
  ci/consistency_gt/official_generated' || 'diffusion-ci/consistency_gt/sglang_generated') }}

# 新增 compute-official-gt-matrix job（仅当启用官方模式时运行）
jobs:
  compute-official-gt-matrix:
    if: github.repository == 'sgl-project/sglang' && inputs.run_official_cases
    runs-on: ubuntu-latest
    outputs:
      matrix: ${{ steps.compute.outputs.matrix }}
      case-count: ${{ steps.compute.outputs.case-count }}
    steps:
      - name: Compute official case matrix
        id: compute
        env:
          OFFICIAL_CASE_IDS: ${{ inputs.official_case_ids }}
```

```
  OFFICIAL_SOURCE_GROUP: ${{ inputs.official_source_group || 'all' }}
# 内联 Python 脚本根据组和 ID 筛选案例，并标记需要 H200 的案例
run: |
  python3 - <<'PY'
  import json, os
  # 定义案例组
  groups = {
    "diffusers": ["flux_2_image_t2i", "flux_2_klein_image_t2i", ...],
    "wan21": ["wan2_1_t2v_1.3b"],
    "ltx23": ["ltx_2.3_two_stage_t2v_2gpus", "ltx_2.3_one_stage_ti2v"],
  }
  # 根据 source_group 筛选组
  # 根据 case_ids 筛选具体案例
  # 将大型案例标记为 h200 runner
  # 输出 JSON 矩阵
  PY
```

评论区精华

无有效讨论；仅有一条 Gemini Code Assist bot 的评论，表示无法为涉及的文件类型生成 review。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，仅影响 CI 工作流。可能的风险包括：
  - H200 runner 条件判断错误导致 job 失败；
  - artifact 过滤规则过严导致必要文件缺失；
  - 并发组键变更可能影响并行工作流的取消行为。但这些风险在已有验证步骤中已被覆盖。
  - 影响：影响限于 diffusion-ci-gt-gen.yml 工作流的使用者。启用官方模式后，GT 生成会切换到官方案例，输出目录不同，且大型案例使用 H200。对核心服务无影响，团队成员需要了解新模式以修改触发配置。
  - 风险标记：CI 流程变更，缺少测试覆盖

关联脉络

- 暂无明显关联 PR