

PR #33931 完整报告

sgl-project/sglang

[diffusion] Exercise the default SP exchange in Z-Image CI

合并时间: 2026-08-07 14:20

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33931>

执行摘要

- 一句话: Z-Image CI 移除显式 Ulysses, 覆盖默认 SP 路径
- 推荐动作: 不值得精读, 但建议在阅读 PR #33923 (zimage 注意力路由 USPAttention) 时一并查看本 PR, 理解默认并行路径覆盖意图。关注点: 确认自动策略选择在 Z-Image 上实际会落到默认 SP 交换而非 fallback 到其他并行; 后续若默认策略变化, 这两个用例可作哨兵。

功能与动机

PR body 明确提出“let the existing 2-GPU Z-Image cases use automatic parallel-strategy selection”, 目的是让默认 SP 交换路径进入 CI 回归。此前

python/sglang/multimodal_gen/test/server/gpu_cases.py 中两个 Z-Image 用例显式固定 ulysses_degree=2, 绕过了默认并行策略选择; 删除后可在 2-GPU 环境下持续验证默认并行路径, 同时保留方形用例作为 SP2 性能守卫、非方形用例作为 padding 与分片兼容守卫, 并继续保留显式 Ulysses 与 Ring 用例的后端兼容覆盖。

实现拆解

1. 变更入口: python/sglang/multimodal_gen/test/server/gpu_cases.py, 该文件维护多 GPU diffusion 服务端启动用例列表。
2. 核心动作: 删除 zimage_image_t2i_2_gpus 与 zimage_image_t2i_2_gpus_non_square 两个 DiffusionTestCase 中 DiffusionServerArgs 的 ulysses_degree=2 参数。ulysses_degree 显式指定 Ulysses 序列并行维度, 删除后由服务端默认并行策略选择逻辑决定, 从而在 2-GPU 环境执行默认 SP 交换。
3. 设计意图: 依据 PR body 的四点说明, 该改动将两个用例的职责从“显式 Ulysses 覆盖”调整为“默认路径守卫”, 与其他显式用例形成互补, 覆盖不丢失。
4. 配套影响: 无新增用例、无新增请求, CI 用例数与模型启动次数均不变; 标签含 run-ci, 意图是让常规 CI 持续跑通该默认路径。无源码、schema 或部署配套改动。

关键文件:

- python/sglang/multimodal_gen/test/server/gpu_cases.py (模块 CI 用例; 类别 test; 类型 test-coverage): 唯一变更文件: 删除两个 Z-Image 2-GPU 用例的显式 ulysses_degree=2, 使其走自动并行策略选择, 在 CI 中覆盖默认 SP 交换路径。

关键符号: 未识别

关键源码片段

python/sglang/multimodal_gen/test/server/gpu_cases.py

唯一变更文件：删除两个 Z-Image 2-GPU 用例的显式 `ulysses_degree=2`，使其走自动并行策略选择，在 CI 中覆盖默认 SP 交换路径。

```
# python/sglang/multimodal_gen/test/server/gpu_cases.py (节选)
# Z-Image 2-GPU 用例：不再显式设置 ulysses_degree=2，改由自动并行策略选择，
# 以便在 CI 中覆盖默认 SP 交换路径（配合 PR #33923 的 USPAttention 路由）。
DiffusionTestCase(
    "zimage_image_t2i_2_gpus",
    DiffusionServerArgs(
        model_path=DEFAULT_SMALL_MODEL_NAME_FOR_TEST,
        # 删除前此处为 ulysses_degree=2，删除后自动选择并行方案
    ),
),

# 非方形用例作为 padding 与分布式分片兼容守卫，同样移除显式 Ulysses 配置，
# 保留 1280x720 输出尺寸，继续验证非方形分片路径。
DiffusionTestCase(
    "zimage_image_t2i_2_gpus_non_square",
    DiffusionServerArgs(
        model_path=DEFAULT_SMALL_MODEL_NAME_FOR_TEST,
        # 删除前此处为 ulysses_degree=2，删除后自动选择并行方案
    ),
    DiffusionSamplingParams(
        prompt=T2I_sampling_params.prompt,
        output_size="1280x720",
    ),
    run_perf_check=False,
),
```

评论区精华

本 PR 没有任何 review 评论或 review 线程。唯一的 issue 评论是作者触发的 /tag-and-rerun-ci，属于 CI 重跑指令，无技术交锋。

- 暂无高价值评论线程

风险与影响

- 风险：行为风险较低：删除显式 `ulysses_degree=2` 后，两个用例依赖自动并行策略选择；若默认策略在 Z-Image 上未实际走到 SP 交换路径，覆盖目标会落空，但不会造成功能错误；若默认路径存在 bug，两个用例可能开始失败，这恰是本 PR 想要的哨兵效果。覆盖偏移风险较低：非方形用例原先在显式 Ulysses 下承担 padding/ 分片兼容守卫，切换后该场景可能变为默认 SP 路径，但仓库中 `wan2_1_i2v_14b_lora_2gpu`、`qwen_image_t2i_2_gpus` (Ring) 等用例仍显式覆盖 Ulysses/Ring，兼容覆盖不丢失。性能与安全无影响，CI 用例数与模型启动次数不变。

- 影响：影响范围仅限 diffusion 多 GPU CI 用例配置，对用户功能与生产代码零影响。收益点是默认 SP 交换路径在 Z-Image 上获得持续回归覆盖，降低未来并行策略默认值变更时的隐形回归风险。CI 用例数量、模型启动次数与请求内容均不变。
- 风险标记：CI 配置变更，依赖默认 SP 路径

关联脉络

- PR #33923 [Diffusion] Route zimage and hunyuanvideo attention through USPAttention: zimage 注意力后端已切到 USPAttention，本 PR 让 Z-Image CI 从显式 Ulysses 转向默认路径，正是为了验证该路由在默认并行策略下的行为。
- PR #32667 [Diffusion] Add K/V-gather sequence parallel attention: 为 diffusion 新增 K/V-gather 序列并行注意力并默认在 SP2 启用，自动策略选择的默认路径很可能就是 K/V-gather SP，本 PR 提供了持续回归入口。
- PR #33707 Derive H3 attention admission from backend capabilities: 并行后端准入改为按能力自证、显式冲突 fail closed，与本 PR 共同体现“显式配置 → 自动选择”的演进主线。