PR #22371 完整报告

sgl-project/sglang

Fix image (random multimodal) dataset token statistics

合并时间：2026-05-17 14:12

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22371

执行摘要

- 一句话：随机多模态数据集 token 统计增强
- 推荐动作：值得精读：展示了如何为调试目的添加细粒度统计信息，不引入风险。适合作为基准测试可观测性增强的参考模式。

功能与动机

当前 SGLang 图像数据集存在 Token 化往返（初始 token → decode → re-encode）不确定性问题，导致实际 Token 数与 random-input-len 指定值不符。此外基准测试只报告总输入 Token 数，无法区分文本、Chat 模板、Token 化往返开销或视觉模板的贡献。PR 旨在输出每个请求的 Token 分解，使用户能直接看出 Token 来源。

实现拆解

1. 提取 DatasetRow 字段：在 sample_image_requests 函数数据集生成循环结束后，从 dataset 中提取 r.text_prompt_len 和 r.vision_prompt_len，并计算文本提示开销（text_prompt_len - input_lens）。
2. 打印 Token 分解表：将上述数据整理为 stat_fields 列表，遍历打印每个维度的平均值和总和，包括原始文本 Token、Chat 模板后 Token、文本提示开销、视觉 Token。
3. 保持原有统计不变：原有总输入 / 输出 Token 数和图像计数打印保留，新增分解表放在两者之间。 涉及的唯一文件：python/sglang/benchmark/datasets/image.py，仅添加约 14 行日志代码。

关键文件：

- python/sglang/benchmark/datasets/image.py（模块 基准测试；类别 source；类型 core-logic）：唯一变更文件，在 sample_image_requests 函数末尾添加 Token 分解统计打印，增强基准测试数据集的可观测性。

关键符号：未识别

关键源码片段

python/sglang/benchmark/datasets/image.py

唯一变更文件，在 sample_image_requests 函数末尾添加 Token 分解统计打印，增强基准测试数据集的可观测性。

```
# 在循环结束后，从 dataset 中提取字段
# text_prompt_len 是应用 chat template 后的文本 token 数
# vision_prompt_len 是视觉 token 数
# input_lens 是原始请求的文本 token 数
# 开销 = chat template 后长度 - 原始长度
text_prompt_lens = np.array([r.text_prompt_len for r in dataset])
vision_prompt_lens = np.array([r.vision_prompt_len for r in dataset])
text_prompt_overheads = text_prompt_lens - input_lens

# 定义要展示的四个维度
stat_fields = [
    ("Raw text prompt tokens (without overhead)", input_lens),
    ("Text prompt tokens (with chat template)", text_prompt_lens),
    ("Text prompt overhead", text_prompt_overheads),
    ("Vision tokens", vision_prompt_lens),
]

print("\n=== Token Breakdown (per request avg / total) ===")
for label, vals in stat_fields:
    # 输出每个维度的平均值和总和，便于对比
    print(f"  {label}: avg={np.mean(vals):.1f}, total={np.sum(vals)}")
```

评论区精华

Reviewer rkarhila-amd 确认变更仅影响随机图像数据集基准测试，并表示视觉模型在 gfx950 GPU 上支持不稳，但该 PR 本身无问题。最终获得 HaiShaw 批准合并。无争议性讨论。

- 视觉模型支持不稳定 (other): 无行动项，PR 本身正常。

风险与影响

- 风险：风险极低。仅添加信息性日志输出，不修改任何推理路径、数据流或状态。新代码仅在基准测试数据集生成时执行，不干扰模型服务。
- 影响：直接影响：使用 sample_image_requests 生成随机多模态数据集的基准测试用户将看到详细的 Token 分解表，辅助分析 Token 数量偏差。其他用户无影响。团队无需额外部署或配置变更。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR