PR #23714 完整报告

sgl-project/sglang

[diffusion] CI: update ground truth with official output

合并时间：2026-04-29 13:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23714

执行摘要

- 一句话：引入官方 GT 优先级并修复对齐
- 推荐动作：建议关注 GT 回退策略设计（优先级 + 存在性检查），可作为跨版本测试基础设施的参考。对于 Qwen-Image 相关开发者，需了解后处理扩展模式，便于其他模型复用。

功能与动机

官方 ground truth 提供了更稳定的基准，减少因 SGLang 实现偏移导致的误报。同时修复部分实现细节以对齐官方输出，提升 CI 可靠性。

实现拆解

1. 重新设计 GT 加载优先级：在 test_utils.py 中新增两个基础 URL 常量，get_consistency_gt_remote_files 和 load_consistency_gt 改为先调用 _find_remote_consistency_gt_files 遍历官方、SGLang、旧目录，按优先级返回第一个匹配文件。
2. Qwen-Image 后处理扩展：在 qwen_image.py 中为 qwen_image_postprocess_text 添加 return_attention_mask 参数，当为 True 时返回 (prompt_embeds, mask) 元组；在 postprocess_cfg_noise 中引入 cfg_scale 变量，当 true_cfg_scale 为 None 时使用 guidance_scale 作为 fallback，避免 TypeError。
3. 文本编码阶段适配：在 text_encoding.py 中，当后处理函数签名包含 return_attention_mask 时传递该参数，并处理返回值元组解包，将有后处理的 mask 优先用于存储。
4. Qwen-Image DiT 注意力掩码传递：在 qwen_image.py (dits) 的 joint attention forward 中从 kwargs 提取 attn_mask 传入 USPAttention，并在 DiT forward 中将 encoder_hidden_states_mask 与图像 mask 拼接后传入各 block。
5. 阈值更新与配置调整：在 consistency_threshold.json 中更新 10+ 个 diffusion 用例的 CLIP/SSIM/PSNR/mean_abs_diff 阈值；在 qwen_image_layered.py 中修复空 prompt 回退到自动 caption 逻辑，修改 mask 属性名以对齐下游。
6. 其他修复：在 component_loader.py 中修正 tokenizer 加载的 use_fast 判断逻辑，仅在架构以 Tokenizer 结尾且不以 Fast 结尾时使用 slow；在 qwen2_5vl.py 中将 mask 创建参数从位置参数改回关键字参数以提高可维护性。

关键文件：

- python/sglang/multimodal_gen/test/test_utils.py（模块 测试工具；类别 test；类型 test-coverage；符号 _remote_consistency_gt_candidates, _remote_file_exists, _find_remote_consistency_gt_files, get_consistency_gt_remote_files）：核心变更：实现官方 GT 优先级加载逻辑，新增远程文件存在性检查函数
- python/sglang/multimodal_gen/configs/pipeline_configs/qwen_image.py（模块 Qwen-Image 配置；类别 source；类型 core-logic；符号 qwen_image_postprocess_text, postprocess_cfg_noise）：语义修复核心：后处理函数返回 mask，CFG noise 使用 fallback 值
- python/sglang/multimodal_gen/runtime/pipelines_core/stages/text_encoding.py（模块 文本编码；类别 source；类型 core-logic）：适配后处理返回 mask，处理 stage 内 mask 存储
- python/sglang/multimodal_gen/runtime/models/dits/qwen_image.py（模块 Qwen-Image 模型；类别 source；类型 data-contract）：传递注意力掩码到 joint attention，支持 encoder_hidden_states_mask
- python/sglang/multimodal_gen/test/server/consistency_threshold.json（模块 阈值配置；类别 test；类型 test-coverage）：调整阈值以匹配官方 GT 输出

关键符号：qwen_image_postprocess_text, postprocess_cfg_noise, get_consistency_gt_remote_files, _find_remote_consistency_gt_files, _remote_file_exists, encode_text, QwenImageJointAttention.forward, QwenImageDiT.forward, QwenImageLayeredPipelineConfig.postprocess_cfg_noise

关键源码片段

python/sglang/multimodal_gen/test/test_utils.py

核心变更：实现官方 GT 优先级加载逻辑，新增远程文件存在性检查函数

```
def _find_remote_consistency_gt_files(
    case_id: str,
    num_gpus: int,
    is_video: bool,
    output_format: str | None = None,
) -> list[tuple[str, str]]:
    # 按优先级遍历所有一致性 GT 基础地址
    for base_url in SGL_TEST_FILES_CONSISTENCY_GT_BASES:
        # 为当前基础地址生成候选文件名
        candidates = _remote_consistency_gt_candidates(
            base_url, case_id, num_gpus, is_video, output_format
        )
        if is_video:
            # 视频时需要三帧全部存在才认为有效
            if all(_remote_file_exists(url) for _, url in candidates):
                return candidates
        else:
            # 图片时返回第一个存在的文件
            for filename, url in candidates:
```

```
            if _remote_file_exists(url):
                return [(filename, url)]
    return []
```

python/sglang/multimodal_gen/configs/pipeline_configs/qwen_image.py

语义修复核心：后处理函数返回 mask，CFG noise 使用 fallback 值

```
def postprocess_cfg_noise(
    self,
    batch,
    noise_pred: torch.Tensor,
    noise_pred_cond: torch.Tensor,
) -> torch.Tensor:
    # Qwen-Image 遵循官方 diffusers 的 true-CFG 行为：
    # 使用 true_cfg_scale 合并 cond/uncond 后，对噪声预测进行逐 token 范数匹配
    # 当 true_cfg_scale 未设置时，使用 guidance_scale 作为 fallback
    cfg_scale = (
        batch.true_cfg_scale
        if batch.true_cfg_scale is not None
        else batch.guidance_scale
    )
    # 如果 cfg_scale 无效或未启用 CFG，直接返回原始预测
    if (
        cfg_scale is None
        or cfg_scale <= 1.0
        or not batch.do_classifier_free_guidance
    ):
        return noise_pred
    # 计算条件分支范数，对噪声预测进行缩放
    cond_norm = torch.norm(noise_pred_cond, dim=-1, keepdim=True)
    noise_norm = torch.norm(noise_pred, dim=-1, keepdim=True).clamp_min(1e-12)
    return noise_pred * (cond_norm / noise_norm)
```

python/sglang/multimodal_gen/runtime/pipelines_core/stages/text_encoding.py

适配后处理返回 mask，处理 stage 内 mask 存储

```
            # ......
            if "return_attention_mask" in postprocess_sig.parameters:
                postprocess_kwargs["return_attention_mask"] = return_attention_mask
            prompt_embeds = postprocess_func(outputs, text_inputs, **postprocess_kwargs)
            has_postprocessed_attention_mask = False
            postprocessed_attention_mask = None
            # 如果后处理函数返回了额外注意力掩码，解包并标记
            if isinstance(prompt_embeds, tuple):
                prompt_embeds, postprocessed_attention_mask = prompt_embeds
                has_postprocessed_attention_mask = True
            # ......
            if return_attention_mask:
```

```
if has_postprocessed_attention_mask:
    mask_to_store = (
        postprocessed_attention_mask.to(device=target_device)
        if postprocessed_attention_mask is not None
        else None
    )
elif attention_mask is not None:
    mask_to_store = attention_mask.to(device=target_device)
else:
    mask_to_store = torch.ones(
        input_ids.shape[:2], device=target_device
    )
attn_masks_list.append(mask_to_store)
```

评论区精华

- gemini-code-assist[bot] 指出 postprocess_cfg_noise 中 cfg_scale 可能同时为 None 导致 TypeError，建议增加 None 检查（已采纳）。
- 同样机器人建议将 use_fast=False 逻辑限制到必要模型，以免影响性能（已调整）。
- 建议保持 mask_kwargs 使用关键字参数，避免位置参数脆弱（已修改）。
- 建议简化空 prompt 判断为 not prompt or prompt.isspace()（已应用）。
- postprocess_cfg_noise 中 cfg_scale 可能为 None (correctness): 建议添加 None 检查
- use_fast=False 的默认逻辑可能影响性能 (performance): 建议更精确的条件判断或添加注释说明
- mask_kwargs 改为位置参数降低可维护性 (design): 建议保持关键字参数
- 简化空 prompt 检查 (style): 建议简化

风险与影响

- 风险：
  - GT 查找依赖远程 URL，网络问题可能导致 CI 失败，但已有 fallback 机制。
  - use_fast 改动可能影响 tokenizer 加载速度，需基准测试。
  - Threshold 调整可能掩盖微小回归，需要结合人工审查。
  - mask 数据流变更可能影响其他模型正确性，需确保回归测试通过。
- 影响：
  - 对用户：无直接影响。
  - 对 CI 系统：一致性检查更可靠，减少因 SGLang 生成差异导致的假阳性。
  - 对团队：维护 GT 需定期同步官方输出，已提供 repro 脚本。
  - 风险标记：GT 远程依赖，tokenizer 性能影响，mask 数据流变更，阈值变更可能掩蔽回归

关联脉络

- 暂无明显关联 PR