PR #22360 完整报告

sgl-project/sglang

[diffusion] fix: fix loading multiple ckpts with different precision for a same module

合并时间：2026-04-09 02:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22360

PR #22360 分析报告

执行摘要

该 PR 修复了扩散模型中因加载多个精度变体检查点导致的不一致性 bug，通过过滤重复文件和添加快速失败检查，确保权重加载确定性，提升 Sana 等模型的稳定性，属于重要的扩散模块维护性修复。

功能与动机

为什么做：根据 PR body，动机是解决 Sana（及其他如 mova 模型）的一致性波动问题。根本原因是目录中同时存在多个精度变体的 safetensors 文件（例如 foo.safetensors 和 foo.fp16.safetensors），加载时重复的参数名称导致最终权重依赖于文件加载顺序，从而引发非确定性行为。

实现拆解

按模块拆解改动：

1. transformer_load_utils.py：新增 _filter_duplicate_precision_variant_safetensors 函数，使用正则表达式 _PRECISION_VARIANT_SUFFIX_RE 识别精度后缀（如 .fp16），当存在非精度变体的规范文件时，过滤掉重复变体。python canonical_path = f"{match.group('stem')}{match.group('shard') or ''}{match.group('ext')}" if canonical_path in canonical_paths: removed.append(path)
2. weight_utils.py：新增 _raise_if_duplicate_safetensors_keys 函数，在迭代 safetensors 文件时检测重复 tensor 键，并抛出 ValueError 快速失败，避免加载顺序依赖。
3. 测试文件：更新 consistency_threshold.json 中多个模型的阈值（如 qwen_image_t2i_cache_dit_enabled 的 CLIP 阈值从 0.92 提升至 0.99），反映修复后稳定性提升；新增 test_transformer_quant.py 单元测试，覆盖过滤逻辑。

评论区精华

review 讨论中的关键交锋：仅有一个 review 评论，由 gemini-code-assist[bot] 提出，但话题偏离核心变更。评论指出 scheduler 的 warmup 过程在多节点环境中会因临时目录路径不共享而失败，建议使用固定路径或共享存储方案。然而，此评论针对的文件 scheduler.py 未在本 PR 中修改，可能是一个无关反馈或误关联。

> 引用评论要点："The use of tempfile.mkdtemp() on the source rank followed by broadcasting the absolute path to all other ranks will cause failures in multi-node deployments."

风险与影响

具体风险：

- 过滤误判：正则表达式可能错误匹配文件路径，导致必要精度变体被过滤，影响模型精度或加载失败。
- 快速失败过度：_raise_if_duplicate_safetensors_keys 函数可能因误报重复键而中断合法加载场景，尤其是在复杂分片配置下。

影响评估：

- 用户影响：修复后，扩散模型（如 Sana、Wan）的输出更稳定，减少性能波动。
- 系统影响：增强加载模块的鲁棒性，降低因文件打包错误导致的不确定风险。
- 团队影响：为类似权重加载问题提供了设计参考，例如优先规范文件和防御性检查。

关联脉络

与历史 PR 的关系：

- PR #21817：同为扩散模块修复，针对 warmup 图像初始化的秩安全问题，共享一致性和加载主题。
- PR #22127：涉及扩散模型的 NVFP4 量化测试，反映团队在扩散和量化领域的持续投入。
  演进趋势：近期多个 PR（如 #21861、#21610）关注性能优化和内核扩展，而本 PR 侧重核心加载路径的确定性修复，表明仓库在功能扩展的同时，也加强基础模块的稳定性和可靠性。