PR #20430 完整报告

sgl-project/sglang

[diffusion][CI] Add CI for MOVA model inference

合并时间：2026-03-25 02:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20430

执行摘要

此 PR 为 MOVA-360p 视频生成模型添加了 CI 测试，通过定义模型常量和配置多个测试用例（单 GPU 和双 GPU），扩展了多模态生成测试套件的覆盖范围。变更简单且风险低，但 review 中讨论了代码组织和重用性问题，建议关注测试配置的设计模式。

功能与动机

PR 的主要动机是将 MOVA-360p 视频生成模型集成到现有测试套件中，以验证其在不同 GPU 配置下的功能正确性。根据 PR body 描述，目标是 'add support for testing the MOVA-360p video generation model'，这有助于确保新模型的稳定性和兼容性，为持续集成提供基础。

实现拆解

实现集中在两个文件：

- test_utils.py：添加了常量 DEFAULT_MOVA_360P_MODEL_NAME_FOR_TEST = "OpenMOSS-Team/MOVA-360p"，为测试提供模型路径。
- testcase_configs.py：将常量添加到可用模型列表，并创建四个测试用例：
  - 单 GPU 案例："mova_360p_1gpu"，使用 DiffusionServerArgs 配置单 GPU 和 dit_layerwise_offload=True。
  - 双 GPU 案例：三个案例分别配置 tp_size=2、ring_degree=1, ulysses_degree=2 和 ring_degree=2, ulysses_degree=1，展示不同并行策略。 所有案例均使用 TI2V_sampling_params 并设置 run_perf_check=False。

评论区精华

review 讨论主要围绕代码优化：

- 设计权衡：mickqian 提问 'could we reuse the existing sampling_params?'，CloudRipple 回复已修复，最终改用 TI2V_sampling_params，避免了冗余定义。
- 代码风格：gemini-code-assist[bot] 建议 'add a comment above this line to group this new model constant' 并减少配置重复，前者已采纳（常量前添加了注释），后者未明确解决，提示未来维护风险。

风险与影响

风险：测试用例中 DiffusionServerArgs 配置存在重复（如 review 指出），可能增加未来更新成本；run_perf_check=False 意味着未验证性能回归，但这是有意设置以减少 CI 开销。 影响：直接影响测试套件，无用户端变化；新增 CI 测试可能轻微延长运行时间，但提升了模型覆盖；团队需注意测试配置的标准化，以保持代码整洁。

关联脉络

从历史 PR 看，此 PR 与 #21042（修改相同 testcase_configs.py 文件）和 #20352（添加 Hunyuan3D 支持）相关，共同显示仓库在扩散模型测试和多模态生成能力上的持续扩展。近期 PR 如 #21041（修复 FLUX.1 模型）也涉及扩散测试修复，表明这是一个活跃的演进领域。