PR #26963 完整报告

sgl-project/sglang

[diffusion] Add Cosmos3 Nano T2V GPU test

合并时间：2026-06-03 15:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26963

执行摘要

- 一句话：添加 Cosmos3 Nano T2V 单 GPU 一致性测试
- 推荐动作：该 PR 属于常规测试补充，逻辑清晰，无争议，建议合并。若有后续 Cosmos3 系列优化 PR，应同步更新对应的测试基线与阈值。

功能与动机

PR body 说明需要为 Cosmos3 Nano T2V 模型增加一个轻量级一致性测试，以覆盖该模型在 CI 中的正确性验证，并固定对应的 gt 数据集（ci-data PR #11）。

实现拆解

1. 在 gpu_cases.py 中添加测试用例：在 T2V 用例列表末尾插入一个新的 DiffusionTestCase，名称为 "cosmos3_nano_t2v"，指定模型路径 DEFAULT_COSMOS3_NANO_MODEL_NAME_FOR_TEST、模态 video、环境变量 SGLANG_DISABLE_COSMOS3_GUARDRAILS=1，采样参数包括 prompt、输出尺寸、帧数、推理步数等，并开启一致性检查、关闭性能检查与组件精度检查。
2. 在 perf_baselines.json 中补充性能基线：为 cosmos3_nano_t2v 添加了各阶段耗时、每步去噪时间、预期端到端时间等占位数据，同时补充了 estimated_full_test_time_s 字段供调度器估算测试时长。此外还补填了其他几个已有用例缺失的 estimated_full_test_time_s 字段（如 zimage_image_t2i_2_gpus、qwen_image_edit_ti2i 等），并新增了 cosmos3_nano_t2i 及 flux_2_t2i_customized_vae_path 等用例的基线。
3. 在 consistency_threshold.json 中添加一致性阈值：新增 "cosmos3_nano_t2v" 键，配置 clip 阈值 0.90、ssim 阈值 0.89、psnr 阈值 24.0、mean_abs_diff 阈值 10.0。

关键文件：

- python/sglang/multimodal_gen/test/server/gpu_cases.py（模块 GPU 用例；类别 test；类型 test-coverage）：新增 cosmos3_nano_t2v 测试用例定义，是本次变更的核心入口。
- python/sglang/multimodal_gen/test/server/perf_baselines.json（模块 性能基线；类别 test；类型 test-coverage）：添加 Cosmos3 Nano T2V 的性能基线（各阶段耗时、去噪步时、预期端到端时间），供调度与监控使用。
- python/sglang/multimodal_gen/test/server/consistency_threshold.json（模块 一致性阈值；类别 test；类型 test-coverage）：设定 Cosmos3 Nano T2V 一致性检查的阈值，保证 CI 中的输出质量门禁。

关键符号：未识别

评论区精华

该 PR 没有 review 评论。仅包含一条自动消息提示 daily quota 耗尽，以及作者触发的 /tag-and-rerun-ci 命令。未发现实质性技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：变更仅涉及测试配置文件（JSON/Python），不触及任何核心推理或运行时逻辑。主要风险在于性能基线值和一致性阈值可能不准确，若环境差异导致测试误报，需后续微调阈值。此外，测试依赖外部 ci-data 仓库的 gt 数据，若 gt 数据更新或丢失可能导致测试失败。
- 影响：影响范围仅限于 diffusion 测试套件中的 Cosmos3 Nano T2V 模型，新增一个轻量级一致性测试（约 4 步推理），预计增加约 65 秒的 CI 耗时。对用户无直接影响，对团队可提升该模型的回归覆盖信心。
- 风险标记：测试依赖外部 ci-data 数据集，阈值可能需随环境微调

关联脉络

- PR #27084 [diffusion] Optimize Cosmos3 i2v latent prep: 同为 Cosmos3 模型相关，优化了 I2V 的潜变量预处理，本 PR 为其提供测试覆盖。