PR #20707 完整报告

sgl-project/sglang

[diffusion] model: support two stage pipeline of LTX-2

合并时间：2026-04-04 09:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20707

执行摘要

本 PR 实现了 LTX-2 扩散模型的两阶段视频生成管道，通过新增上采样器、扩展管道阶段和更新配置，支持低分辨率生成后上采样精炼。变更涉及文档、核心管道、模型加载和测试，提升了视频生成质量，是扩散模块的重要功能扩展。

功能与动机

PR 旨在支持 Lightricks/LTX-2 模型的两阶段生成功能，以生成更高分辨率的视频。从 CLI 示例可见，用户需指定 `--pipeline-class-name LTX2TwoStagePipeline` 来启用此功能，同时可覆盖空间上采样器和蒸馏 LoRA 路径。动机是集成该先进模型到 SGLang 框架，增强视频生成能力，满足用户对高质量输出的需求。

实现拆解

- 文档更新：docs/diffusion/compatibility_matrix.md 添加 LTX-2 支持，说明两阶段用法和自动路径解析。
- 配置修改：python/sglang/multimodal_gen/configs/pipeline_configs/ltx_2.py 调整管道参数，如 VAE 压缩比和 generator_device 默认值。
- 新增上采样器：python/sglang/multimodal_gen/runtime/models/upsampler/latent_upsampler.py 实现空间上采样模型，包含 BlurDownsample 和 PixelShuffleND 等组件。
- 管道阶段扩展：python/sglang/multimodal_gen/runtime/pipelines_core/stages/upsampling.py 新增 LTX2HalveResolutionStage、LTX2LoRASwitchStage 和 LTX2UpsampleStage，处理分辨率调整、LoRA 切换和上采样逻辑。
- 注册逻辑：python/sglang/multimodal_gen/registry.py 新增 has_registered_diffusion_model_path 函数，改进模型检测以支持 LTX-2。
- CLI 和服务器参数：python/sglang/cli/utils.py 和 python/sglang/multimodal_gen/runtime/server_args.py 扩展参数解析，支持 --pipeline-class-name 和组件路径（如 --spatial-upsampler-path）。
- 测试基准：python/sglang/multimodal_gen/test/server/perf_baselines.json 添加 ltx_2_two_stage_t2v 性能基准用例。

评论区精华

Review 评论为空，仅由 mickqian 批准。提交历史显示 60 次提交，包括多次调试和修复，如对齐潜变量布局、修复 CFG 路径和调整 LoRA 合并，表明实现细节经过迭代优化，但讨论未在 review 中公开记录。

风险与影响

- 技术风险：新增上采样器可能影响推理性能；管道复杂性增加维护难度；自动路径解析可能失败于网络或文件问题；LoRA 切换机制可能引入不稳定性；测试覆盖有限，仅有一个性能基准用例。
- 影响分析：对用户提供新功能，提升视频生成体验；系统代码库复杂度增加，需持续维护新组件；团队需熟悉扩散模型的两阶段生成逻辑，可能增加学习成本。

关联脉络

与近期 PR 如 #22040（扩散模型 CLI 修复）同属扩散模块演进，显示团队在扩展模型支持和优化用户体验。历史 PR 中多涉及性能优化（如 JIT 激活回滚）和测试改进，本 PR 延续了功能增强趋势，是扩散模型支持的重要一步。