PR #23366 完整报告

sgl-project/sglang

[diffusion] model: support LTX2.3 high quality pipeline

合并时间：2026-04-24 14:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23366

执行摘要

- 一句话：添加 LTX-2.3 高质量两阶段生成流水线
- 推荐动作：建议精读 ltx_2_denoising.py 中的 res2s 采样器实现（特别是 _ltx2_res2s_sde_step 和 _ltx2_stage2_res2s_step），这是与官方 HQ 对齐的核心算法；同时关注 _merge_lora_into_data 的分组合并优化，可推广到其他 LoRA 场景。HQ 的 sigma 调度和重噪声生成器设计也值得参考。

功能与动机

LTX-2.3 官方提供了高质量两阶段流水线（TI2VidTwoStagesHQPipeline），SGLang 原先仅支持基础流水线，无法达到官方 HQ 的输出质量。PR body 中给出了 SpongeBob 示例的 PSNR 对比（20.71 dB），并附上了完整的官方复现代码，表明需要对齐官方 HQ 行为。

实现拆解

1. 新增 HQ 采样参数与 CLI 解析

- configs/sample/ltx_2.py新增 LTX23HQSamplingParams 类，继承 LTX23SamplingParams，覆盖默认分辨率（1088×1920）、推理步数（15）、蒸馏 LoRA 强度等，并通过 build_request_extra 将 distilled_lora_strength_stage_1/2 注入请求扩展字段。
- entrypoints/cli/generate.py新增 _resolve_cli_sampling_params_cls 函数，根据 pipeline_class_name 或模型信息动态选择采样参数类，使 CLI 能正确解析 HQ 专属参数。

2. 实现 HQ 去噪核心（res2s 采样器与时间步缩放）

- pipelines_core/stages/ltx_2_denoising.py在 LTX2DenoisingStage 中添加完整的 res2s RK2 中点 SDE 采样器，包括：
  - _ltx2_channelwise_normalize：对噪声张量进行逐通道标准化。
  - _ltx2_res2s_new_noise / _ltx2_res2s_noise_like：基于固定种子的确定性噪声生成（使用 float64 高精度）。
  - _ltx2_phi_1 / _ltx2_phi_2 / _ltx2_get_res2s_coefficients：用于中点更新的指数积分器。
  - _ltx2_get_sde_coeff / _ltx2_res2s_sde_step：带 SDE 噪声的 RK2 步骤。
  - _ltx2_stage2_res2s_step：两阶段 res2s 整合（速度→x0→中点重估→最终组合）。 同时修改 __init__ 接受 sampler_name 参数以支持 euler/res2s 切换。

- models/dits/ltx_2.py新增 _scale_timestep_for_adaln 方法，在 HQ 模式下将时间步乘以 timestep_scale_multiplier（1000），对齐官方 AdaLN 输入语义。所有 AdaLN 调用（包括 prompt 和 cross-attention）均经过此缩放。

3. 适配两阶段流水线（sigma 准备、LoRA 强度、重噪声生成）

- pipelines/ltx_2_pipeline.py：
  - LTX2SigmaPreparationStage.forward：对于 HQ 流水线，根据视频 latent token 数量计算分辨率感知的 sigma 移位（number_of_tokens 参数），而非固定锚点。
  - _add_ltx2_stage1_generation_stages 增加 denoising_sampler_name 参数，透传给去噪阶段。
  - 新增 switch_lora_phase、_get_stage_distilled_lora_strength、_can_short_circuit_lora_switch 等方法，支持按阶段配置蒸馏 LoRA 强度（stage1 默认为 0.0，stage2 默认为 1.0），并允许请求级覆盖。
  - LTX2TwoStagePipeline（基础类）增强参数快照机制，当快照缺失时自动克隆当前张量而非抛异常。
- pipelines_core/stages/denoising_av.py：
  - LTX2RefinementStage 新增 _build_stage2_renoise_generator 和 _ltx2_renoise_like：为 HQ 流水线创建与官方对齐的确定性重噪声生成器（基于请求种子并前进 stage-1 的 packed 形状），替代原先的 batch.generator 自然前进。
  - forward 中新增 HQ 分支：使用 fp32 中间计算、sigma 调度末尾追加 [0.0011, 0.0] 以匹配官方最后一步行为。
- pipelines_core/stages/latent_preparation_av.py：记录 ltx2_stage1_packed_video_shape 和 ltx2_stage1_packed_audio_shape 到 batch.extra，供重噪声生成器前进种子使用。

4. 优化 LoRA 合并性能与兼容性

- layers/lora/linear.py：_merge_lora_into_data 重构为分块合并（LORA_MERGE_CHUNK_BYTES = 32MB），避免超大 lora_B 导致显存溢出；同时支持 lora_B_sliced 为非 Tensor 对象（如 quantized），并正确应用 add_ 的 alpha 缩放。
- pipelines_core/stages/text_connector.py：修复 attention mask 计算，使用 torch.finfo(dtype).max 替代硬编码 -1000000.0，避免 HQ 场景下值域溢出。

5. 测试与文档配套

- test/server/gpu_cases.py新增 HQ 测试用例，覆盖两阶段 T2V/TI2V 场景。
- test/server/perf_baselines.json添加 HQ 性能基线。
- docs_new/docs/sglang-diffusion/compatibility_matrix.mdx和 docs/diffusion/compatibility_matrix.md更新兼容性矩阵。
- registry.py注册 LTX2TwoStageHQPipeline 的配置类和采样参数类。

关键文件：

- python/sglang/multimodal_gen/runtime/pipelines_core/stages/ltx_2_denoising.py（模块去噪阶段；类别 source；类型 core-logic；符号 _randn_like_with_batch_generators,

_ltx2_channelwise_normalize, _ltx2_res2s_new_noise,_ltx2_init_res2s_noise_generators）：核心去噪阶段，新增完整 res2s 采样器（RK2 中点SDE），包括噪声生成、通道标准化、指数积分器及 midpoint re-eval 路径；同时修改初始化支持采样器选择。是 HQ 流水线算法核心。

- python/sglang/multimodal_gen/runtime/pipelines/ltx_2_pipeline.py（模块 流水线；类别 source；类型 dependency-wiring；符号 _add_ltx2_stage1_generation_stages, _can_short_circuit_lora_switch, _get_stage_distilled_lora_strength, switch_lora_phase）：HQ 流水线类与 sigma 调度核心；修改 LTX2SigmaPreparationStage 支持分辨率感知 sigma 移位，新增可配置蒸馏 LoRA 强度逻辑，增强参数快照机制。
- python/sglang/multimodal_gen/runtime/pipelines_core/stages/denoising_av.py（模块 去噪阶段；类别 source；类型 core-logic；符号 _build_stage2_renoise_generator, _ltx2_renoise_like）：阶段 2 重噪声生成器与 sigma 调度扩展；新增 _build_stage2_renoise_generator 实现确定性重噪声种子前进，HQ 分支在 sigma 末尾追加 0.0011 以匹配官方最后一步行为。
- python/sglang/multimodal_gen/runtime/models/dits/ltx_2.py（模块 模型层；类别 source；类型 data-contract；符号 _scale_timestep_for_adaln）：DiT 模型新增 _scale_timestep_for_adaln，在 HQ 模式下将时间步乘以 1000 对齐官方 AdaLN 语义；所有 AdaLN 调用（含 prompt/cross-attention）均经过缩放。
- python/sglang/multimodal_gen/configs/sample/ltx_2.py（模块 采样参数；类别 source；类型 core-logic；符号 LTX23HQSamplingParams, build_request_extra）：新增 LTX23HQSamplingParams 类，定义 HQ 默认参数（分辨率 1088x1920、步数 15、蒸馏强度）并注入请求扩展字段。
- python/sglang/multimodal_gen/runtime/layers/lora/linear.py（模块 LoRA 层；类别 source；类型 core-logic）：LoRA 合并重构为分块合并，避免大 lora_B 张量显存溢出；支持非 Tensor lora_B_sliced 并正确应用 add_ 缩放。

关键符号：_add_ltx2_stage1_generation_stages, _can_short_circuit_lora_switch, _get_stage_distilled_lora_strength, _build_stage2_renoise_generator, _ltx2_renoise_like, _scale_timestep_for_adaln, LTX23HQSamplingParams.build_request_extra, _resolve_cli_sampling_params_cls, _ltx2_res2s_new_noise, _ltx2_stage2_res2s_step

关键源码片段

python/sglang/multimodal_gen/runtime/pipelines_core/stages/ltx_2_denoising.py

核心去噪阶段，新增完整 res2s 采样器（RK2 中点 SDE），包括噪声生成、通道标准化、指数积分器及 midpoint re-eval 路径；同时修改初始化支持采样器选择。是 HQ 流水线算法核心。

```
# LTX-2.3 HQ 使用的确定性噪声生成与通道标准化
# 在 float64 下生成标准正态噪声，先全局标准化再逐通道标准化
@classmethod
def _ltx2_res2s_new_noise(
    cls,
    reference_tensor: torch.Tensor,
```

```
    generator: torch.Generator,
) -> torch.Tensor:
    # float64 高精度噪声，确保与官方 repo 的 numpy 采样一致
    noise = torch.randn(
        reference_tensor.shape,
        generator=generator,
        dtype=torch.float64,
        device=reference_tensor.device,
    )
    # 全局标准化
    noise = (noise - noise.mean()) / noise.std()
    # 逐通道标准化 (C, H, W) 维度
    return noise.sub_(noise.mean(dim=(-2, -1), keepdim=True)).div_(
        noise.std(dim=(-2, -1), keepdim=True)
    )

# 初始化两个固定种子的噪声生成器（步骤级和子步骤级）
@staticmethod
def _ltx2_init_res2s_noise_generators(ctx: LTX2DenoisingContext) -> None:
    reference_tensor = (
        ctx.latents if isinstance(ctx.latents, torch.Tensor) else ctx.audio_latents
    )
    if reference_tensor is None:
        raise ValueError("LTX-2 res2s requires video or audio latents.")
    device = reference_tensor.device
    ctx.res2s_step_noise_generator = torch.Generator(device=device).manual_seed(
        LTX23_RES2S_STEP_NOISE_SEED # -1
    )
    ctx.res2s_substep_noise_generator = torch.Generator(device=device).manual_seed(
        LTX23_RES2S_SUBSTEP_NOISE_SEED # 9999
    )
```

评论区精华

- 审核者 nidhishgajjar 要求完善 PR 描述：该 PR 最初为草稿模板，但合并前已填充完整描述并附有官方复现代码。
- gemini-code-assist 指出缩进问题：在 ltx_2_denoising.py 中，step.current_model 调用参数缩进不一致（c95cde6）。PR 作者在后续提交中已修复该问题。
- PR 描述完整性 (other)：最终合并前已填充完整描述和官方复现代码，但提交历史中无明确对应修改记录。
- 缩进对齐问题 (style)：已在后续提交中修复。

风险与影响

- 风险：
  - 遗留路径回归风险：HQ 新增的 sigma 移位、重噪声生成、时间步缩放等均通过 pipeline_class_name == "LTX2TwoStageHQPipeline" 门控，但提交历史显示曾有回归

（如 b7869e8 修复 sigma 移位无意影响遗留路径）。需确保 CI 覆盖全部 LTX-2.3 变体。

- LoRA 合并数值精度：_merge_lora_into_data 重构为分组合并与 add_ 缩放，若 lora_B_sliced 为非 Tensor 类型，可能引入数值差异。现有测试未覆盖全部量化场景。
- 确定性噪声种子变更：HQ 专用重噪声生成器（_build_stage2_renoise_generator）改变了 stage-2 的随机行为，若用户依赖旧版种子序列，输出可能变更。但官方 HQ 要求确定性行为。
- 性能影响：分组合并在 CPU 端增加循环开销，但对 GPU 并行无影响；HQ 的 res2s 采样器比 euler 多一次模型调用（中点重估），计算量约增加 30%。

- 影响：
  - 用户影响：LTX-2.3 用户可通过指定 --pipeline-class-name LTX2TwoStageHQPipeline 启用 HQ 流水线，获得与官方一致的输出质量。现有 LTX-2.3 用户（使用基础流水线）不受影响。
  - 系统影响：新增约 1500 行代码，关键路径增加分支判断；LoRA 合并性能优化（分块）对所有使用 LoRA 的模型均有正面影响。
  - 团队影响：需要维护两条流水线路径（基础 /HQ），新增采样器策略（res2s）增加了去噪阶段的复杂度。
  - 风险标记：遗留路径回归，LoRA 数值精度，噪声种子变更，分块合并性能

关联脉络

- 暂无明显关联 PR