

PR #31491 完整报告

sgl-project/sglang

Feat/spectrum

合并时间: 2026-08-06 04:23

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31491>

执行摘要

- 一句话: 新增 Spectrum 跳步加速, Flux/Wan/Hunyuan 去噪提速 3-5x
- 推荐动作: 值得精读。设计亮点包括: Chebyshev 岭回归 + 局部 Taylor 混合预测 (w 参数)、无 LAPACK/MAGMA 构建下的纯 Python Cholesky 回退、get_total_forward_steps 对 CFG 分支计数语义的精确建模、以及 "显式门控避免污染编译图" 的工程纪律。评审中 alexnails 对 SD3 真 CFG 双分支污染的深度分析 (含数据流推导) 尤其值得学习。建议阅读时重点关注 spectrum.py 的 forecaster 生命周期与 4 个模型 forward 的门控模式, 可作为后续接入新 DiT 模型的参照。

功能与动机

PR body 明确说明动机: "This PR adds Spectrum-based diffusion acceleration to reduce denoising latency while preserving output quality, and includes safety/validation improvements to make the feature robust for production use."。即在不训练、不显著掉画质的前提下压缩扩散采样延迟, 并配套生产级护栏: 拒绝 enable_teacache 与 enable_spectrum 同时开启、让 Spectrum 调度逻辑仅在启用时运行、补充预测器行为单测与平滑信号误差回归守卫。

实现拆解

实现拆解如下:

1. 预测器核心 (新增 runtime/cache/spectrum.py, 约 645 行): 实现 ChebyshevForecaster (Chebyshev 基岭回归, 滑动窗口缓存 (step, 展平特征))、SpectrumForecaster (Chebyshev 与局部离散 Taylor 预测的混合, w 控制混合权重)、SpectrumMixin (挂在 CachableDiT 上的调度生命周期: begin_spectrum_step / spectrum_record_features / spectrum_predict_features) 与 SpectrumContext (当前 step、总 forward 次数、CFG 状态)。岭回归求解 _ridge_cholesky_solve 优先走 torch.linalg.cholesky, 数值病态时加 jitter 重试, 遇到无 LAPACK/MAGMA 的 PyTorch 构建 (如部分 ROCm wheel) 则回退到纯 Python 实现的 Cholesky 分解与求解。拟合过程统一在 fp32 下进行 (autocast 会把 matmul 降成 bf16, 破坏 torch.cholesky_solve 的 dtype 约束), 输出再恢复为原始特征 dtype。评审后改为预分配 (K, N) 环形缓冲区, 避免每步 cat 重排。
2. 参数与 CLI 接入 (新增 configs/sample/spectrum.py, 修改 configs/sample/sampling_params.py): SpectrumParams 继承 CacheParams, 提供

window_size、flex_window（即论文 alpha）、warmup_steps、w、lam、m、history_size、tau_num_steps、taylor_order 等参数；`get_total_forward_steps` 根据 `do_cfg` 与 `separate_cfg_branches` 计算单个 Spectrum 计数器在一次生成中看到的 forward 次数（单计数器 + 真 CFG 时为 $2 * \text{num_inference_steps}$ ），同时服务于 tau 归一化与分支计数器回绕。`SamplingParams` 新增 `enable_spectrum / spectrum_params` 字段，在 `__post_init__` 中惰性实例化 `SpectrumParams`，在 `_validate` 中强制 `teacache` 与 `spectrum` 互斥，并在 `add_cli_args / get_cli_args` 中提供 `--enable-spectrum` 与整套 `--spectrum-*` 参数，通过谱标志映射聚合成 `spectrum_params` 字典。当用户提供了 `spectrum` 专属参数但未开 `--enable-spectrum` 时会自动启用并打印一次性日志，避免参数被静默忽略。

3. 模型前向接入（`dits/base.py`、`flux.py`、`stablediffusion3.py`、`hunyuanvideo.py`、`wanvideo.py`）：`CachableDiT` 改为同时继承 `SpectrumMixin` 与 `TeaCacheMixin`，构造时初始化 Spectrum 状态。4 个模型的 forward 在 Transformer 块入口增加门控：仅当 `forward_batch.enable_spectrum` 为真时才调用 `begin_spectrum_step()`；返回 False 表示本步可跳过全部 Transformer 块，用 `spectrum_predict_features(hidden_states)` 的预测特征继续后续 `norm_out` 等；真实 forward 之后调用 `spectrum_record_features(hidden_states)` 收集样本。Wan/Hunyuan/SD3 每个 CFG 分支维护独立计数器，FLUX 共享单计数器（FLUX.1-dev 用内嵌 guidance，实际单分支）。评审修复了两个关键点：FLUX/SD3 增加显式 `enable_spectrum` 门控，避免把 Spectrum 逻辑编入所有用户的编译图；恢复 rebase 时丢失的 `freqs_cis` 与 `num_replicated_prefix` 传参。

4. 测试配套（新增 `test_spectrum.py`，扩展 `test_sampling_params.py`）：覆盖 Chebyshev 拟合 / 预测的 shape 与 dtype、bf16 与 CUDA autocast 下的行为、w 混合路径、平滑信号低误差回归守卫（相对 L2 误差 $< 1e-3$ ）、tau 归一化与参考实现一致（固定 50 步 horizon），以及 `teacache/spectrum` 互斥校验与 CLI 标志映射。均为离线单元测试，不依赖真实模型权重与 GPU。

关键文件：

- `python/sglang/multimodal_gen/runtime/cache/spectrum.py`（模块 缓存模块；类别 source；类型 core-logic；符号 `_flatten`, `_unflatten`, `_is_missing_linalg_backend_error`, `_cholesky_lower`）：新增 645 行核心实现：`ChebyshevForecaster / SpectrumForecaster / SpectrumMixin / SpectrumContext`，含 LAPACK/MAGMA 缺失时的纯 Python Cholesky 回退与预分配环形缓冲。
- `python/sglang/multimodal_gen/configs/sample/spectrum.py`（模块 采样参数；类别 source；类型 core-logic；符号 `SpectrumParams`, `get_total_forward_steps`）：`SpectrumParams` 参数模型与 `get_total_forward_steps` 的 CFG 分支计数语义，是跳步调度的核心决策点。
- `python/sglang/multimodal_gen/configs/sample/sampling_params.py`（模块 采样参数；类别 source；类型 configuration；符号 `SamplingParams`, `post_init`, `_validate`, `add_cli_args`）：`enable_spectrum` 开关、`SpectrumParams` 惰性实例化、`teacache` 互斥校验、全套 CLI 参数与自动启用逻辑。
- `python/sglang/multimodal_gen/runtime/models/dits/base.py`（模块 模型前向；类别 source；类型 data-contract；符号 `CachableDiT`）：`CachableDiT` 混入 `SpectrumMixin`,

把 Spectrum 与 TeaCache 并列进统一的 DiT 缓存加速基类。

- `python/sglang/multimodal_gen/runtime/models/dits/flux.py` (模块 模型前向; 类别 source; 类型 core-logic; 符号 FluxTransformer2DModel) : FLUX forward 增加跳步门控与特征记录 / 预测; 评审中发现并修复了 rebase 时丢失的 `freqs_cis / num_replicated_prefix` 传参。
- `python/sglang/multimodal_gen/runtime/models/dits/stablediffusion3.py` (模块 模型前向; 类别 source; 类型 core-logic; 符号 SD3Transformer2DModel) : SD3 forward 接入 Spectrum 门控; PR 自述 SD3 上游存在回归, 未承诺端到端正确性, 评审中还暴露了真 CFG 双分支污染问题。
- `python/sglang/multimodal_gen/runtime/models/dits/hunyuanvideo.py` (模块 模型前向; 类别 source; 类型 core-logic; 符号 forward) : HunyuanVideo 接入 Spectrum, CFG 双分支独立计数器模式的代表模型之一。
- `python/sglang/multimodal_gen/runtime/models/dits/wanvideo.py` (模块 模型前向; 类别 source; 类型 core-logic; 符号 forward) : WanVideo 接入 Spectrum, 本 PR 基准测试中去噪加速比最高的模型 ($\alpha=3.0$ 时 4.79x) 。
- `python/sglang/multimodal_gen/test/unit/test_spectrum.py` (模块 单元测试; 类别 test; 类型 test-coverage; 符号 TestSpectrumForecaster, test_chebyshev_fit_and_predict, test_chebyshev_fit_and_predict_bfloat16, test_chebyshev_fit_and_predict_bfloat16_under_autocast) : 新增预测器单测: shape/dtype/autocast 行为、平滑信号误差回归守卫 (相对 $L2 < 1e-3$)、tau 归一化与参考实现一致性。
- `python/sglang/multimodal_gen/test/unit/test_sampling_params.py` (模块 单元测试; 类别 test; 类型 test-coverage; 符号 test_teacache_and_spectrum_are_mutually_exclusive, test_get_cli_args_maps_spectrum_prefixed_flags) : 覆盖 teacache/spectrum 互斥校验与 spectrum 前缀 CLI 标志映射。
- `python/sglang/multimodal_gen/runtime/cache/__init__.py` (模块 缓存模块; 类别 source; 类型 dependency-wiring) : 导出 Spectrum 相关符号, 使 runtime/cache 成为统一缓存加速包入口。
- `python/sglang/multimodal_gen/configs/models/dits/stablediffusion3.py` (模块 模型配置; 类别 source; 类型 configuration) : SD3 相关配置微调 (+1 行), 配合模型前向接入。

关键符号: ChebyshevForecaster.update, ChebyshevForecaster.predict, ChebyshevForecaster._fit_if_needed, _ridge_cholesky_solve, SpectrumForecaster.predict, SpectrumParams.get_total_forward_steps, SamplingParams.get_cli_args, begin_spectrum_step, spectrum_record_features, spectrum_predict_features

关键源码片段

`python/sglang/multimodal_gen/runtime/models/dits/flux.py`

FLUX forward 增加跳步门控与特征记录 / 预测; 评审中发现并修复了 rebase 时丢失的 `freqs_cis / num_replicated_prefix` 传参。

```
# 显式判断 enable_spectrum 后才把调度逻辑编入图 (review 反馈: 不能无条件调用,
```

```

# 否则 Spectrum 会进入每一个 FLUX 用户的编译图)
forward_batch = get_forward_context().forward_batch
spectrum_enabled = forward_batch is not None and forward_batch.enable_spectrum
run_transformer_blocks = self.begin_spectrum_step() if spectrum_enabled else True

if run_transformer_blocks:
    # 真实 forward 路径：双流块与单流块保持原有调用方式,
    # freqs_cis 与 num_replicated_prefix 是文本 / 图像分片的关键参数, 遗漏会导致错误
    for block in self.transformer_blocks:
        encoder_hidden_states, hidden_states = block(
            hidden_states=hidden_states,
            encoder_hidden_states=encoder_hidden_states,
            temb=temb,
            freqs_cis=freqs_cis,
            joint_attention_kwargs=joint_attention_kwargs,
            num_replicated_prefix=num_replicated_prefix,
        )
    for block in self.single_transformer_blocks:
        encoder_hidden_states, hidden_states = block(
            hidden_states=hidden_states,
            encoder_hidden_states=encoder_hidden_states,
            temb=temb,
            freqs_cis=singles_freqs_cis,
            joint_attention_kwargs=joint_attention_kwargs,
            num_replicated_prefix=num_replicated_prefix,
        )
    if spectrum_enabled:
        # 真实结果写入滑动窗口, 后续跳步依赖这批特征做岭回归拟合
        self.spectrum_record_features(hidden_states)
    else:
        if spectrum_enabled:
            # 跳步: 用 Chebyshev 外推特征替换真实 forward 结果, 跳过全部 transformer 块
            hidden_states = self.spectrum_predict_features(hidden_states)

```

评论区精华

评审中产生了多轮高质量技术交锋，核心议题如下：

- SD3 真 CFG 下单计数器双分支污染：alexnaills 分享了 Claude 的深度分析——单计数器模型在真 CFG 下 cond/uncond 共用同一 denoising step 横坐标，ridge 拟合实际回归到两分支平均特征，且 Taylor 路径因 $dt \approx 0$ 被 clamp 到 $1e-8$ 后放大 k 到约 $1e8$ 产生垃圾输出。jaron1990 最初以 "SD3 fails on the main sglang branch" 反驳，后在复测中确认 "seems like claude wasn't wrong and I fixed it"（提交 ff05bd2）。
- 无条件调用 `begin_spectrum_step` 污染编译图：alexnaills 指出 "This is called unconditionally, which puts Spectrum inside the compiled graph for every FLUX user, even when not enabled", 要求显式门控；jaron1990 在 FLUX 与 SD3 中补上 `enable_spectrum` 判断。

- 每步 `cat` 重分配导致内存膨胀: alexnails 建议改为增量维护 XtX/XtH 或预分配环形缓冲, jaron1990 采纳后者 (提交 7df60ce) 。
- end-of-run 清空 forecaster 的取舍: jaron1990 消融后发现清空会破坏最终跳步路径、导致输出质量退化, 因此保留 " 仅在生成开始时 reset" 的生命周期策略, alexnails 认可并致谢。
- CLI 自动启用 Spectrum: alexnails 质疑 `--spectrum-*` 参数未配 `--enable-spectrum` 时自动启用是否为预期行为; jaron1990 确认是刻意设计 (避免参数被静默忽略), 并补充一次性日志。
 - SD3 真 CFG 下单计数器双分支污染 (correctness): 确认为真实缺陷并修复, 但 " 单计数器 + 真 CFG" 组合仍缺少专项测试覆盖。
 - begin_spectrum_step 无条件调用污染编译图 (performance): jaron1990 在 FLUX 与 SD3 中补上 enable_spectrum 显式 gate (提交 3860573) 。
 - 每步 cat 重分配导致内存膨胀 (performance): jaron1990 采用预分配环形缓冲方案 (提交 7df60ce) 。
 - FLUX forward 丢失 freqs_cis 与 num_replicated_prefix (correctness): 提交 78c9997 恢复参数, 确认模型前向改动的敏感性与 rebase 风险。
 - end-of-run 是否清空 forecaster (design): 保留 " 仅在生成开始时 reset" 的生命周期策略, alexnails 认可并致谢 (提交 70eb4a0) 。
 - spectrum 参数自动启用与 CLI 命名 (design): jaron1990 确认自动启用是刻意设计 (避免参数被静默忽略), 并补充一次性日志 (提交 009e2b0) 。
 - fp32 反复转换与 tau scale 预计算 (performance): jaron1990 将 H 缓冲改为直接以 fp32 存储 (提交 2d56318), 并在 __init__ 中预计算 _tau_scale (提交 04ec6a6) 。
 - w=1.0 时浪费 Taylor 计算 (performance): jaron1990 采纳按 w 分支短路, 避免默认配置下的多余计算。

风险与影响

- 风险:
 - 正确性风险 (核心前向路径) : Flux、SD3、Wan、Hunyuan 4 个模型的 forward 都有改动, 虽然以 enable_spectrum 条件分支保护、默认关闭时走原路径, 但任何分支遗漏 (如 FLUX 曾丢失 freqs_cis / num_replicated_prefix) 都会破坏现有生成。SD3 真 CFG 的双分支污染问题虽已修复, 但 " 单计数器 + 真 CFG" 组合仍缺少专项测试。
 - 内存风险: ChebyshevForecaster 默认 history_size=100, 环形缓冲预分配 (100, N) 的 fp32 张量, N 为展平后的 DiT 隐状态维度。Flux 1024² 分辨率下 N 可达千万级, 单实例可能占用数 GB 显存, 需按模型特征维度评估 K 值。
 - 质量风险: 跳步以质量换速度, flex_window 越大越快但质量越差。目前只有合成平滑信号的回归测试, 缺少真实图像的端到端质量回归, 视频模型 (Wan/Hunyuan) 尤其缺乏视觉对比验证。
 - 生命周期风险: forecaster 只在生成开始时 reset, 跨请求保留状态; 曾有清空尝试因质量退化回退, 说明状态复用行为敏感, 需要更明确的隔离策略。

- 已知限制：PR 自述 SD3 因上游 SGLang 回归当前不保证端到端正确性，Spectrum hooks 仅作完整性保留。
- 影响：
 - 用户侧：diffusion 用户可通过 `--enable-spectrum` 获得显著去噪加速（Flux denoise 3.16x-4.30x、Wan 3.43x-4.79x、Hunyuan 3.44x-4.80x，数据来自 PR body 与复审测试），默认关闭不影响现有用户；质量 / 速度由 `flex_window` 与 `w` 参数调节。
 - 系统侧：新增与 TeaCache 并列、互斥的第三类缓存加速（`cache_type="spectrum"`），开启时占用显存存储特征环形缓冲。
 - 团队侧：为 diffusion 加速引入 training-free 特征预测范式，SpectrumMixin + CachableDiT 的组合可作为后续更多模型接入的模板，也为与 TeaCache 的对比 / 融合留下演进空间。
- SD3 用户：当前不承诺 SD3 端到端正确性，使用前需关注上游回归修复状态。
- 风险标记：核心前向路径变更（Flux/SD3/Wan/Hunyuan），环形缓冲区内内存占用（ $K \times N$ 展平特征），SD3 真 CFG 双分支曾污染预测器，缺少视频模型端到端质量回归，SD3 上游回归未承诺端到端正确性

关联脉络

- PR #33546 [diffusion] Wan VAE RMSNorm+SiLU fusion behind quality=high (H200 FastWan2.2 e2e 9.611 -> 9.125 s): 同属 diffusion 端到端延迟优化，且引入了 quality 门控来管理速度 / 质量权衡，与 Spectrum 的 `flex_window` 参数设计同属一条演进方向；均修改 `multimodal_gen` 运行时路径。
- PR #33703 [diffusion] Add SageAttention packed varlen path for MiniMax-H3: 同为 diffusion 去噪阶段性能优化，关注模型特定 forward 路径与后端选择，与 Spectrum 在模型前向中做条件分支的改法同构。
- PR #30683 [Diffusion] Batch GLM-Image AR requests: 同为 diffusion 管线级性能优化（吞吐 / 延迟），引入了 `dynamic_batch_admission` 等调度机制，显示 `multimodal_gen` 性能优化是持续演进方向。