PR #20706 完整报告

sgl-project/sglang

[diffusion] Unify `TeaCacheParams` and `WanTeaCacheParams`

合并时间：2026-03-28 09:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20706

执行摘要

本次 PR 统一了 TeaCache 参数类，将 WanTeaCacheParams 集成到标准 TeaCacheParams 中，通过添加动态系数回调和步骤跳过边界，简化不同扩散模型的 teacache 处理，减少代码重复并提高可扩展性。

功能与动机

根据 PR 描述，动机是“统一 WanTeaCacheParams 到标准 TeaCacheParams 类，使处理不同模型的 teacache 更容易”。这是一个独立的变更，源于 #19957，旨在消除 Wan 模型的特殊参数类，提供更通用的系数选择方式，以支持动态配置和简化维护。

实现拆解

主要改动点如下：

模块	关键变更	影响
参数配置(teacache.py)	扩展 TeaCacheParams 类：添加 coefficients_callback、start_skipping、end_skipping 字段；新增 get_coefficients 方法；移除 WanTeaCacheParams 类。	统一参数接口，支持回调驱动系数选择。
模型适配(wan.py)	定义回调函数 _wan_1_3b_coefficients 和 _wan_14b_coefficients；更新 Wan 采样参数类使用 TeaCacheParams，设置 coefficients_callback。	Wan 模型现在通过回调动态选择系数，减少硬编码。

模块	关键变更	影响
运行时逻辑(wanvideo.py)	修改 should_skip_forward_for_cached_states 方法：使用 get_skip_boundaries 处理 CFG 逻辑，简化步骤边界计算；移除对 WanTeaCacheParams 的检查。	优化 CFG 模式下的 teacache 跳过行为，提高代码清晰度。
辅助修改	- sampling_params.py: 扩展 _json_safe 函数支持 callable 序列化。	
- test_sampling_params.py: 添加单元测试验证回调优先级和 Wan 边界兼容性。	确保参数可 JSON 化，并提供回归测试覆盖。	

评论区精华

review 中，主要讨论集中在 wanvideo.py 的 CFG 处理逻辑：

- adarshxs：询问“could you explain the logic here?”，关注新实现细节。
- eitanturok：解释旧代码如何预加倍 ret_steps 和 cutoff_steps 造成混淆，新实现通过 get_skip_boundaries 方法直接处理 CFG 模式，简化了逻辑。引用原话：“When CFG is enabled, the transformer's forward() is called twice per denoising step... The old code handled this by pre-doubling... which was confusing.”
- 结论：adarshxs 表示理解，讨论已解决，PR 获得 mickqian 批准。

风险与影响

- 风险：
  - 回调序列化可能失败，但通过 _json_safe 函数返回模块限定名缓解。
  - 步骤边界计算变更可能导致 teacache 跳过行为回归，但单元测试 test_wan_teacache_boundaries_match_legacy_behavior 验证了与旧行为的一致性。
  - 系数选择逻辑从硬编码改为回调，可能引入运行时类型错误，但测试覆盖了回调优先级。
- 影响：
  - 用户：Teacache 参数使用更一致，Wan 模型行为应保持兼容，不影响现有功能。
  - 系统：减少代码冗余，提高维护性；新回调机制为未来模型扩展提供基础。
  - 团队：简化 CFG 逻辑，降低未来开发复杂度。

关联脉络

此 PR 是 SGLang 中 diffusion 模块持续演进的一部分。关联 PR 包括：

- 21600（支持覆盖层模型）：扩展扩散模型功能，与本 PR 的参数统一相辅相成。
- 21407（修复 Flux2-Klein 标记化）：涉及配置调整和测试覆盖，类似本 PR 的测试增强。
- 20633（清理冗余预处理函数）：同为代码重构，提升代码质量。 这些 PR 共同推动扩散组件向更模块化、可维护的方向发展，本 PR 通过统一参数类为这一趋势添砖加瓦。