PR #21190 完整报告

sgl-project/sglang

[Whisper] Enable CUDA graph support and timestamp for whisper model

合并时间：2026-03-29 01:46

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21190

执行摘要

该 PR 为 Whisper 模型启用了 CUDA 图支持和时间戳功能，通过重构交叉注意力实现，移除了不兼容的缓存机制，实现了 36% 的吞吐量提升，同时保持精度不变，并扩展了 API 以支持时间戳输出，显著提升了语音转录服务的性能和功能。

功能与动机

之前，Whisper 模型使用自定义的 bmm + mask 交叉注意力实现和 Python 端的 _encoder_cache 字典，这不兼容 CUDA 图捕获 / 重放，导致无法利用图形优化提升性能。该 PR 旨在解决此问题，启用 CUDA 图以提升推理效率（相关 Issue #21161），并添加时间戳支持以兼容 OpenAI 的 verbose_json 格式，增强用户体验。PR body 中强调目标是实现 "36% throughput improvement" 且 "identical accuracy"。

实现拆解

实现按模块拆解如下：

- 模型层(python/sglang/srt/models/whisper.py)：将手动 BMM 交叉注意力替换为 RadixAttention 路径，代码从约 91 行缩减到 32 行，简化逻辑并启用 CUDA 图。关键变更：forward 方法中，交叉注意力时设置 k=None, v=None，依赖 self.attn 处理缓存。
- 执行器层(python/sglang/srt/model_executor/cuda_graph_runner.py 和 model_runner.py)：修复 CUDA 图捕获，设置 encoder_len_fill_value 为 max_source_positions（如 Whisper 的 1500），确保交叉注意力内核被包含在捕获图中。
- 注意力后端(python/sglang/srt/layers/attention/flashinfer_backend.py)：修复 update_cross_attention 函数，使用 encoder_lens 而非 seq_lens 进行交叉注意力规划，避免 KV 长度覆盖错误。
- API 和处理器：
  - python/sglang/srt/entrypoints/openai/serving_transcription.py：新增 _parse_segments 函数解析时间戳令牌，支持 verbose_json 响应。
  - python/sglang/srt/multimodal/processors/whisper.py：修改 _pop_sampling_param 处理时间戳粒度，调整解码器提示令牌（使用 <|0.00|> 替代 <|notimestamps|>）。
  - python/sglang/srt/entrypoints/http_server.py：扩展 API 端点支持 timestamp_granularities 和 verbose_json 格式。
- 配置和测试：

- python/sglang/srt/server_args.py：自动为 Whisper 选择 flashinfer 后端，并禁用 radix cache 以避免前缀缓存冲突。
- test/manual/test_whisper_cuda_graph.py：新增测试文件，验证 CUDA 图支持的正确性和请求一致性。

评论区精华

Review 评论中仅有一人（mickqian）批准，无具体技术讨论。Issue 评论中主要有两个线程：

- Lint修复：yuan-luo评论"Pleasefixlint."，作者通过多次运行CI（如/tag-and-rerun-ci）解决格式问题，体现了团队对代码质量的关注。
- 性能基准测试：作者提供了与 vLLM 的详细基准测试对比，显示 SGLang 在 Whisper 服务上的性能优势，但无进一步技术交锋，结论是变更被接受。

风险与影响

风险：

1. 回归风险：whisper.py 中交叉注意力逻辑变更可能引入错误，但新增测试覆盖和精度验证（WER 不变） mitigates 此风险。
2. 兼容性问题：server_args.py 中自动选择 flashinfer 后端可能影响其他编码器 - 解码器模型，需确保后端支持充分测试。
3. 时间戳解析稳定性：_parse_segments 函数需处理异常令牌序列，边缘情况（如缺失时间戳令牌）可能导致解析失败或输出错误。
4. CUDA 图捕获依赖：修复依赖于 max_source_positions 配置，若模型配置错误（如非 Whisper 模型），可能导致捕获失败或性能退化。

影响：

- 用户：获得显著的性能提升（吞吐量 +36%）和新增时间戳功能，提升语音转录体验。
- 系统：CUDA 图启用减少推理延迟，提升资源利用率，但增加了后端复杂性（如强制使用 flashinfer）。
- 团队：代码变更跨多个模块，增加了维护负担，但通过测试和文档（PR body 中的基准）提供了清晰的质量保证。

关联脉络

从历史 PR 分析看，该 PR 与以下 PR 相关：

- PR 20441：修复 Piecewise CUDA Graph 崩溃，同是 CUDA 图优化，显示仓库对图形捕获技术的持续改进。
- PR 21123：减少多模态张量哈希的 CPU 内存，共享性能优化主题，体现团队在多模态和推理效率上的演进方向。 整体上，该 PR 是 sglang 仓库在提升多模型（如 Whisper）推理性能的重要一步，结合了架构重构（RadixAttention 集成）和新功能扩展（时间戳），为后续类似编码器 - 解码器模型的优化提供了参考模板。