PR #22089 完整报告

sgl-project/sglang

[Feature] Add chunk-based streaming ASR for Qwen3-ASR

合并时间：2026-04-10 01:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22089

执行摘要

本 PR 为 SGLang 的 Qwen3-ASR 模型添加了服务器端块流式转录功能，通过将音频分割为 2 秒块并应用前缀回退算法，实现了实时部分转录输出。该变更显著降低了首次文本延迟，但以编码重复计算为代价，为多模态流式处理奠定了基础。

功能与动机

基于 Issue #22025 的流式输入需求，旨在减少用户等待完整音频处理的时间。PR body 引用 Qwen3-ASR 论文算法，目标是通过 SSE 提供逐词输出，提升实时交互体验。

实现拆解

- 流式状态管理：新增 python/sglang/srt/entrypoints/openai/streaming_asr.py，定义 StreamingASRState 类管理前缀回退状态，参数如 chunk_size_sec=2.0。
- 音频分块：split_audio_chunks 函数将音频分割为累积块，每块包含从起始到当前位置的音频。
- 请求处理：在 python/sglang/srt/entrypoints/openai/serving_transcription.py 中，_generate_chunked_asr_stream 方法处理流式请求，每个块作为独立 SGLang 请求发送，检测客户端断开。
- 适配器扩展：适配器接口新增流式支持属性，Qwen3-ASR 适配器实现具体配置，提示模板从处理器导入确保一致性。

评论区精华

- 关键修复：JustinTong0323 指出 StopAsyncIteration 逃逸问题，SammLSH 改用 async for...break 解决。
- 设计权衡：针对提示模板重复，讨论后决定从处理器导入，平衡依赖关系。
- 已知限制：CJK 语言回退无效，因 str.split() 不适用于无空格文本，记录为 TODO。
- 未来方向：API 参数暴露和代码重构被标记为后续任务，保持 PR 最小化。

风险与影响

- 性能风险：编码重复导致线性增长开销，长音频处理效率低。
- 正确性风险：CJK 回退机制失效，可能影响转录质量。

- 兼容性：仅支持 Qwen3-ASR，需适配器扩展以支持其他模型。
- 影响：为用户提供实时转录，但增加系统负载；为团队引入流式框架，促进后续优化。

关联脉络

本 PR 直接依赖 PR #22073，后者添加了 Qwen3-ASR 模型支持。关联 Issue #22025 驱动流式设计，同时参考 vLLM 的流式实现讨论（Issue #35767 和 #35908），显示跨项目的最佳实践对齐。