# PR #22361 完整报告

sgl-project/sglang

[Whisper] Batch encoder forward for concurrent prefill requests

合并时间：2026-04-12 14:15

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22361

## 执行摘要

该 PR 优化了 SGLang 中 Whisper 模型的前向传播，将编码器从串行执行改为批量处理，显著提升了高并发预填充场景下的吞吐量。在 GB300 平台上 64 并发时吞吐量提升达 28.9%，而 B200 平台由于编码器开销较小提升约 3%。修改范围集中，风险低，已通过准确性（WER 未变）和性能基准测试验证。

## 功能与动机

问题背景：当预填充批次包含多个新的 Whisper 请求时，编码器原本在 for 循环中串行运行，每个调用在 B200 上约 7ms、GB300 上约 18ms。这导致调度器被长时间阻塞（例如 5 个请求阻塞 35-90ms），期间无法执行解码批次，成为吞吐量扩展的瓶颈。性能分析显示，编码器在高并发时占调度器时间的 35%。

优化目标：通过批量执行编码器调用，减少调度器阻塞时间，提升高并发下的吞吐量。

## 实现拆解

主要修改位于 python/sglang/srt/models/whisper.py 的 forward 方法：

1. 设备放置修复：cherry-pick 了 PR #22293 的修复，确保输入特征和位置 ID 移动到正确设备： python device = self.conv1.weight.device input_features = input_features.to(device=device) position_ids = position_ids.to(device=device)
2. 特征收集逻辑重构：
   - 将原本的 encoder_list 循环改为收集未缓存特征到 features_to_encode 列表。
   - 跳过已缓存或无效的请求。
3. 批量编码器执行： python if features_to_encode: features_batch = torch.cat(features_to_encode, dim=0) encoder_len = features_batch.shape[-1] // 2 encoder_position_ids = torch.arange(encoder_len, device=features_batch.device) batched_output = self.encoder(features_batch, encoder_position_ids, forward_batch) encoder_hidden_states = batched_output.reshape(-1, batched_output.shape[-1])
4. 输出适配：将批量输出从 [N, seq_len, dim] 重塑为 [N*seq_len, dim]，以适配下游交叉注意力 KV 缓存。

关键设计决策：利用编码器天然的批次兼容性（Conv1d、位置嵌入、32 个 transformer 层、LayerNorm 均支持批次维度），实现无跨请求交互的批量处理，确保逻辑正确性。

## 评论区精华

review 讨论较为简洁，仅有一次实质性建议：

> gemini-code-assist[bot]："For better performance and to avoid an unnecessary CPU-to-GPU transfer, you can create the encoder_position_ids tensor directly on the target device."

该建议被采纳，最终代码中使用了 device=features_batch.device 直接创建张量，避免了额外的设备间传输开销。yhyang201 批准了 PR，无其他争议。

## 风险与影响

技术风险：

- 低风险：编码器本身完全支持批次处理，无跨请求交互，逻辑正确性有保障。
- 设备兼容性：包含 #22293 的设备放置修复，避免设备不匹配问题。
- 准确性：基准测试显示 WER（12.77-12.78%）在所有配置下未变化。
- 内存：批次张量拼接可能轻微增加内存峰值，但音频特征维度固定且批次大小有限，影响可控。

影响范围：

- 性能提升：在 GB300 平台上效果显著（64 并发时吞吐量 +28.9%），B200 平台提升较小（+3%），反映了不同硬件上编码器开销的差异。
- 调度优化：减少调度器阻塞时间，改善整体资源利用率。
- 用户价值：提升语音识别服务的并发处理能力和响应速度。

## 关联脉络

该 PR 与历史 PR #22293 直接相关，后者提供了设备放置修复，被 cherry-pick 到当前优化中作为基础。从近期历史 PR 看，SGLang 仓库持续进行多平台（AMD、NPU、Intel XPU）性能优化和 CI 改进，本 PR 延续了针对特定模型组件（如 Whisper 编码器）的精细化性能调优趋势。同时，标签中的 diffusion、deepseek、npu 反映了该优化可能跨多个模型和平台受益，体现了代码复用和平台扩展的协同演进。