PR #21952 完整报告

sgl-project/sglang

[New Model] Gemma 4

合并时间：2026-04-07 11:24

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21952

执行摘要

此 PR 实现了 Google Gemma 4 模型家族在 SGLang 中的全面支持，涵盖文本、视觉和音频多模态输入，以及推理与工具调用功能。通过新增模型文件、优化 Triton 内核和集成解析器，显著扩展了系统能力。尽管存在一些未决设计问题和兼容性风险，但整体是一个里程碑式的功能增强，值得团队关注和学习。

功能与动机

Gemma 4 是 Google 的新一代开放模型家族，具有 Dense/MoE 架构、多模态支持和原生工具调用能力。PR body 明确指出动机是“添加 Gemma 4 模型支持”，以利用这些先进特性丰富 SGLang 的模型生态，满足用户对多模态 AI 应用的需求。引用 PR 描述：“Gemma 4 is Google's next-generation family of open models featuring Dense and MoE architectures, multimodal support (text, image, audio), hybrid reasoning, and native tool calling.”

实现拆解

实现按模块拆解如下：

- 模型层：新增 gemma4_causal.py（文本模型）、gemma4_mm.py（多模态集成）、gemma4_vision.py（视觉编码器）和 gemma4_audio.py（音频编码器），其中视觉编码器使用 ClippableLinear 进行激活裁剪，音频编码器适配 Conformer 架构。
- 配置系统：在 model_config.py 中注册 Gemma4ForCausalLM 和 Gemma4ForConditionalGeneration 架构，并处理混合 SWA 层（滑动窗口与全注意力）的标识符映射，例如：python elif "Gemma4ForCausalLM" in model_architectures or " Gemma4ForConditionalGeneration" in model_architectures: layer_types = getattr(hf_text_config, "layer_types", []) swa_attention_layer_ids = [i for i, x in enumerate(layer_types) if x == "sliding_attention"]
- 推理与工具调用解析：在 reasoning_parser.py 中添加 Gemma4Detector 类，使用 <|channel> 和 <channel|> 令牌；在 gemma4_detector.py 中实现工具调用解析，支持流式处理和复杂参数格式。
- 性能优化：在 triton_backend.py 中为混合 SWA 层引入 swa_attn_logits 缓冲区以处理不同 v_head_dim；新增 fused kernel gemma_rmsnorm_residual_scalar 融合 RMSNorm、残差加法和标量乘法，减少内核启动。

- 多模态处理：更新 chat_template.py 添加 Gemma-4-it 模板，在 serving_chat.py 中集成推理启用逻辑，并通过 gemma4.py 处理器处理图像 / 音频输入。
- 其他调整：更新代码拼写忽略列表、添加 MoE 调优配置文件、修复 ROCm 兼容性等。

评论区精华

review 讨论中的关键交锋包括：

- gemini-code-assist[bot] 指出设计隐患：在 gemma4_mm.py 第 260 行的 TODO 注释警示 chunked prefill 可能破坏双向注意力，影响图像质量，建议跟踪解决。
- weakref.proxy 移除争议：ispobock 询问移除原因，kpham-sgl 解释为规避 torch.compile 失败（类似历史问题），最终团队同意移除以提升兼容性。
- 测试覆盖验证：ispobock 质疑推理解析器测试缺失，kpham-sgl 确认已补充单元测试，体现了测试驱动开发的重要性。
- 配置细节审核：gemini-code-assist[bot] 提示 MoE 配置文件可能 typo，虽未明确解决，但凸显了性能调优需谨慎。

风险与影响

- 技术风险：核心模型路径变更可能引入回归，特别是在混合 SWA 层处理；依赖特定 transformers 版本（需安装指定提交）带来兼容性挑战；多模态编码器增加系统复杂性，可能影响稳定性和维护负担。
- 影响评估：用户可直接部署 Gemma 4 进行多模态任务，扩展了 SGLang 的应用场景；系统需适配新架构，可能对内存和计算资源有更高要求；团队需更新 CI 和文档，并学习新代码以有效维护。

关联脉络

从历史 PR 看，本 PR 与近期模型支持工作（如 PR #22073 添加 Qwen3-ASR）一脉相承，反映了 SGLang 持续扩展多模态模型生态的趋势。同时，与推测解码（PR #22203）和模型修复（PR #21522）相关，表明团队在功能增强和稳定性优化上并行推进。整体上，这揭示了仓库向更复杂、高性能模型支持演进的战略方向，Gemma 4 的集成是其中的关键一步。