PR #30871 完整报告

sgl-project/sglang

Add VLM prefill profiler ranges

合并时间：2026-07-12 14:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/30871

执行摘要

- 一句话：VLM prefill 添加 profiler 命名范围
- 推荐动作：建议直接合并。该 PR 是低风险、高收益的性能可观测性改进，尤其对 VLM 模型调优有直接帮助。

功能与动机

PR 描述指出，在性能剖析 trace 中，VLM prefill 的视觉 embedding 阶段和 LLM prefill 阶段无法区分。添加命名范围后，可以分别归因两个阶段，例如 Qwen3-VL-32B H100 TP=4 上可测出 mm embedding/ViT 约 7.16 ms，LLM prefill 约 57.16 ms。

实现拆解

1. 添加 sglang.vlm.mm_embedding 范围：将原有的多模态 embedding 获取逻辑（包括 _embed_mm_inputs_with_split 或 embed_mm_inputs 调用）整体包裹在 with torch.profiler.record_function("sglang.vlm.mm_embedding") 中。
2. 添加 sglang.vlm.language_model_prefill 范围：将语言模型 forward 调用 language_model(...) 包裹在 with torch.profiler.record_function("sglang.vlm.language_model_prefill") 中。
3. 无行为改动：所有缩进和逻辑保持不变，仅增加 context manager。

关键文件：

- python/sglang/srt/managers/mm_utils.py（模块 多模态；类别 source；类型 core-logic；符号 general_mm_embed_routine, torch.profiler.record_function）：唯一变更文件，在 VLM prefill 核心函数 general_mm_embed_routine 中为多模态 embedding 和语言模型 prefill 添加了 torch.profiler.record_function 包装。

关键符号：general_mm_embed_routine, torch.profiler.record_function

关键源码片段

python/sglang/srt/managers/mm_utils.py

唯一变更文件，在 VLM prefill 核心函数 general_mm_embed_routine 中为多模态 embedding 和语言模型 prefill 添加了 torch.profiler.record_function 包装。

```
# 文件：python/sglang/srt/managers/mm_utils.py
```

```
# 在 general_mm_embed_routine 函数中，为 VLM 的视觉编码和 LLM prefill 阶段添加 profiling
范围

def general_mm_embed_routine(
    # ... 参数列表保持不变 ...
):
    # ... 前置逻辑保持不变 ...

    if not forward_batch.forward_mode.is_decode() and not forward_batch.forward_mode.is_
    target_verify():
        # ... 数据准备逻辑保持不变 ...

        server_args = get_server_args()
        # Makes VLM profiles directly attributable: this range includes
        # encoder/ViT execution and multimodal feature placement, while
        # the language model range below excludes both.
        with torch.profiler.record_function("sglang.vlm.mm_embedding"): #
        新增：包裹视觉编码阶段
            if server_args and server_args.enable_adaptive_dispatch_to_encoder:
                input_embeds, other_info = _embed_mm_inputs_with_split(
                    # ... 参数不变 ...
                )
            else:
                input_embeds, other_info = embed_mm_inputs(
                    # ... 参数不变 ...
                )
        # ... deepstack 和 offload 逻辑保持不变 ...

    # ... 后续逻辑保持不变 ...

    # Language model prefill 阶段
    with torch.profiler.record_function("sglang.vlm.language_model_prefill"): # 新增：包裹 LLM
    forward
        hidden_states = language_model(
            input_ids=None,
            forward_batch=forward_batch,
            input_embeds=input_embeds,
            **kwargs,
        )

    return hidden_states
```

评论区精华

PR 有两个自动 review 评论：gemini-code-assist[bot] 确认无 review comment，无额外讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅添加 torch.profiler.record_function 上下文管理器，无任何业务逻辑、控制流或数据变更。在非 profiling 模式下开销可忽略。
- 影响：影响范围：仅修改了 VLM 模型推理中的 prefill 路径，且为纯 instrumentation 添加。对端上性能无影响，但为性能分析团队提供了精确的耗时归因能力，有助于优化 VLM 推理瓶颈。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR