PR #30298 完整报告

sgl-project/sglang

[LoRA] Laguna: per-layer LoRA hidden-dim resolution for packed attention

合并时间：2026-08-05 05:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/30298

执行摘要

- 一句话：修复 Laguna 分层头数致 LoRA 生成崩溃
- 推荐动作：值得精读。该 PR 展示了在非均匀 per-layer 注意力几何下如何正确扩展 LoRA 维度解析：提取公共 fallback 函数、在模型类实现按层 hook、并让所有非注意力模块委托回公共逻辑。review 中关于“错误的 fallback 比显式报错更危险”的讨论对错误处理设计有普遍启发；新增的 hermetic 测试（用真实 config 构造 fake model）也是很好的单测实践。后续可关注是否应将 per-layer 维度解析能力下沉到通用层，减少各模型重复实现。

功能与动机

PR body 明确说明：Serving a raw PEFT LoRA adapter on Laguna — targeting q_proj/k_proj/v_proj/o_proj — crashes at generation，报错为 sgemm_lora_a.py 中 assert x.shape[-1] == K。根因是 “Laguna uses per-layer Q-head asymmetry”，“config.num_attention_heads is a single global value (the first full-attention layer's count)”，而通用 fallback 对每一层都使用该全局值，导致头数不同的层上 qkv_proj/o_proj 的 LoRA A/B 缓冲区尺寸错误。

实现拆解

1. 提取公共 helper：在 python/sglang/srt/lora/utils.py 中，将原 get_hidden_dim 的 fallback 分支体（约 100 行）整体抽取为 get_default_hidden_dim(module_name, config, layer_idx, lora_added_vocab_size=0)，get_hidden_dim 在模型未定义 hook 时调用它。此步为纯重构，行为不变，目的是让模型级 hook 有干净的委托目标，避免复制大段分支逻辑。
2. 新增 Laguna 模型 hook：在 python/sglang/srt/models/laguna.py 中为 LagunaModel 添加 get_hidden_dim，对 qkv_proj 与 o_proj 直接使用 config.num_attention_heads_per_layer[layer_idx] 与 config.head_dim 计算维度，其余模块（MLP、MoE、embed、lm_head）委托 get_default_hidden_dim，保证 --lora-target-modules all 仍可用；同时 LagunaForCausalLM.get_hidden_dim 转发到内部 model。此处读取 config.head_dim 而不设 fallback，缺失时显式抛 AttributeError。
3. 新增 hermetic 单元测试：新增 test/registered/unit/lora/test_laguna_hidden_dim_unit.py，基于真实 LagunaConfig 构造 fake model（不初始化权重、纯 CPU），覆盖首层 / 宽层 qkv_proj/o_proj 维度、与全局 fallback 的分歧、非注意力模块委托、未知模块报错、ForCausalLM 转发以及缺失 head_dim 抛 AttributeError 共 10 个用例，并注册

CUDA/AMD CI。

4. 基准验证：在 8×H100 上对 XS.2/XS-2.1/S-2.1 验证修复前崩溃、修复后正常服务；未启用 LoRA 的基础路径延迟对比在噪声范围内，无回归。

关键文件：

- python/sglang/srt/lora/utils.py（模块 LoRA 工具；类别 source；类型 dependency-wiring；符号 get_hidden_dim, get_default_hidden_dim）：提取 get_default_hidden_dim，将通用 fallback 逻辑下沉为公共 helper，是本次重构的核心，为所有模型提供委托目标且不改变既有行为。
- python/sglang/srt/models/laguna.py（模块 模型定义；类别 source；类型 data-contract；符号 get_hidden_dim）：新增 LagunaModel.get_hidden_dim 与 LagunaForCausalLM.get_hidden_dim，以 per-layer 头数解析 LoRA 维度，是修复的核心。
- test/registered/unit/lora/test_laguna_hidden_dim_unit.py（模块 单元测试；类别 test；类型 test-coverage；符号 _make_fake_laguna, TestLagunaPerLayerAttentionDims, test_qkv_proj_uses_first_layer_head_count, test_qkv_proj_uses_wider_layer_head_count）：新增 hermetic CPU 单元测试，覆盖 per-layer 维度解析、与通用 fallback 的分歧、缺失 head_dim 报错、非注意力模块委托与 ForCausalLM 转发，是回归防护。

关键符号：get_default_hidden_dim, LagunaModel.get_hidden_dim, LagunaForCausalLM.get_hidden_dim

关键源码片段

python/sglang/srt/lora/utils.py

提取 get_default_hidden_dim，将通用 fallback 逻辑下沉为公共 helper，是本次重构的核心，为所有模型提供委托目标且不改变既有行为。

```
def get_hidden_dim(module_name, config, base_model, layer_idx, lora_added_vocab_size=0):
    # 模型类可定义 get_hidden_dim 覆盖层相关模块；否则走全局 fallback
    if hasattr(base_model, "get_hidden_dim"):
        return base_model.get_hidden_dim(module_name, layer_idx)
    # 通用 fallback：按均匀注意力几何从 config 推导维度
    return get_default_hidden_dim(module_name, config, layer_idx, lora_added_vocab_size)


def get_default_hidden_dim(module_name, config, layer_idx, lora_added_vocab_size=0):
    """按 config 推导模块的 LoRA 输入 / 输出维度，假设各层注意力几何一致。

    这是模型未定义 ``get_hidden_dim`` 时的 fallback。对于 per-layer 几何
    不一致的模型（如 Laguna 的逐层注意力头数），应在模型类实现
    ``get_hidden_dim``，仅覆盖依赖层的模块，其余模块委托回本函数，
    而不要重新推导所有分支。
    """
    head_dim = getattr(config, "head_dim", config.hidden_size // config.num_attention_heads)
    if module_name == "qkv_proj":
        # 输出维 = head_dim * ( 全局 Q 头数 + KV 头数 * 2)
```

```
            return config.hidden_size, head_dim * (
                config.num_attention_heads + config.num_key_value_heads * 2
            )
        elif module_name == "o_proj":
            o_head_dim = getattr(config, "v_head_dim", None) or head_dim
            # 输入维 = o_head_dim * 全局 Q 头数，输出维 = hidden_size
            return o_head_dim * config.num_attention_heads, config.hidden_size
        # ... 其余分支（gate_up_proj / down_proj / MoE / embed / lm_head 等）
```

python/sglang/srt/models/laguna.py

新增 LagunaModel.get_hidden_dim 与 LagunaForCausalLM.get_hidden_dim，以 per-layer 头数解析 LoRA 维度，是修复的核心。

```
    def get_hidden_dim(self, module_name: str, layer_idx: int) -> Tuple[int, int]:
        """返回模块的 LoRA 输入 / 输出维度，遵循 Laguna 的 per-layer 注意力宽度。

        Laguna 每层注意力由 ``num_attention_heads_per_layer[layer_idx]`` 决定
        （见 ``LagunaAttention``），而 `config.num_attention_heads` 只是全局值，
        对头数不同的层会算错；通用 fallback `get_default_hidden_dim` 就会踩坑，
        导致 `qkv_proj` / `o_proj` 的 LoRA 缓冲区尺寸错误，生成期在
        `sgemm_lora_a.py` 触发 `assert x.shape[-1] == K`。
        这里只覆盖两个注意力投影，其余模块（MLP / MoE / embed / lm_head）委托
        给公共 helper，保证 `--lora-target-modules all` 依然可用。
        """
        config = self.config
        # 不提供 fallback：Laguna 的 head_dim 与 hidden_size // num_attention_heads
        # 不一致（例如 XS.2 为 2048 // 48 = 42，真实值为 128），缺失时宁可抛
        # AttributeError，也不静默算错。
        head_dim = config.head_dim
        num_heads = config.num_attention_heads_per_layer[layer_idx]
        num_kv_heads = config.num_key_value_heads
        if module_name == "qkv_proj":
            return config.hidden_size, head_dim * (num_heads + num_kv_heads * 2)
        elif module_name == "o_proj":
            return head_dim * num_heads, config.hidden_size
        # 非注意力模块（MLP / MoE / embed / lm_head）交给全局 fallback 逻辑
        return get_default_hidden_dim(module_name, config, layer_idx)
```

test/registered/unit/lora/test_laguna_hidden_dim_unit.py

新增 hermetic CPU 单元测试，覆盖 per-layer 维度解析、与通用 fallback 的分歧、缺失 head_dim 报错、非注意力模块委托与 ForCausalLM 转发，是回归防护。

```
def _make_fake_laguna(num_attention_heads_per_layer):
    # 用真实 LagunaConfig 构造模型替身，不跑 __init__（无权重、纯 CPU）
    # head_dim 故意取 128（不等于 hidden_size // 全局头数），确保 hook
    # 必须直接读 config.head_dim 而非推导
    config = LagunaConfig(
        hidden_size=2048,
        head_dim=128,
```

```
        num_key_value_heads=8,
        num_hidden_layers=len(num_attention_heads_per_layer),
        num_attention_heads_per_layer=list(num_attention_heads_per_layer),
    )
    model = LagunaModel.__new__(LagunaModel)
    model.config = config
    return model


def test_wider_layer_differs_from_generic_fallback(self):
    # 回归守卫：宽层（layer_idx=1）的 hook 必须与全局 fallback 结果不同，
    # 否则缓冲区错位会在 sgemm_lora_a.py 触发 assert
    for module_name in ("qkv_proj", "o_proj"):
        hook = self.model.get_hidden_dim(module_name, layer_idx=1)
        generic = get_default_hidden_dim(module_name, self.model.config, 1)
        self.assertNotEqual(hook, generic)
```

评论区精华

核心讨论围绕两点：一是 Jiminator 指出 laguna.py 中 head_dim 的 fallback（hidden_size // num_attention_heads）本身就是错误来源，例如 S-2.1 上会得出 64 而真实 head_dim 为 128，建议直接 config.head_dim 且无 fallback，宁可 AttributeError 也不要静默算错；作者接受并修复，同时新增测试断言缺失时抛错。二是 Jiminator 建议单元测试改用真实 LagunaConfig，避免手工复制 num_attention_heads、first_k_dense_replace 等派生字段导致测试失真，作者照做且断言无需变化。另在 issue 中，Jiminator 补充 XS-2.1 与 XS.2 几何相同（[48,64]），建议加入受影响表格，作者已更新 PR 描述。

- head_dim fallback 推导错误，应移除 (correctness): 作者改为 head_dim = config.head_dim（无 fallback），并新增 test_missing_head_dim_raises_not_derives 用例，缺失时显式抛 AttributeError。
- 单元测试应基于真实 LagunaConfig (testing): 作者重写测试为真实 LagunaConfig，断言无需任何变化。
- XS-2.1 与 XS.2 几何一致，建议补充受影响表格 (question): 作者已在更新后的描述中同时列出 XS.2、XS-2.1 与 S-2.1 三个受影响变体。

风险与影响

- 风险：风险点集中在 lora/utils.py 的公共函数路径：get_hidden_dim 是所有模型的 LoRA 维度入口，提取 get_default_hidden_dim 虽是纯重构，但若未来其它非均匀几何模型误用该 fallback 仍会出错，因此文档明确要求这类模型实现 hook。LagunaModel.get_hidden_dim 直接读取 config.head_dim，缺失时会抛 AttributeError，这是有意为之（fail loudly），但可能影响依赖旧行为的调用方。此外，本次只覆盖 qkv_proj/o_proj 两个 packed projection，若 Laguna 未来新增其它非均匀模块需扩展 hook。测试为 CPU hermetic，未覆盖 GPU 端到端 Triton 路径，但基准已覆盖三种非均匀变体与 null adapter。

- 影响：对用户：使用 Laguna XS.2/XS-2.1/S-2.1 并开启 --enable-lora（target 为 q/k/v/o）的场景从生成崩溃变为可用；M.1/M.1-FP8 等 uniform 变体为 no-op，行为不变。对系统：LoRA 缓冲分配对非均匀头数模型更健壮，get_default_hidden_dim 成为新的公共 API，后续模型可复用。对团队：确立了 per-layer 非均匀模型下 LoRA 维度解析的范式（模型类定义 get_hidden_dim、覆盖层相关模块并委托其余），与 Llama4、qwen3_5 等既有 hook 风格一致。整体影响面中等偏小，默认路径无行为变化。
- 风险标记：核心 LoRA 维度解析路径变更，head_dim 缺失改为显式报错，仅覆盖非均匀头数变体，测试未覆盖 GPU 端到端

关联脉络

- PR #24204 Laguna base support（PR body 提及）：Laguna 模型基础支持 PR，本 PR 依赖其 per-layer 注意力结构与 num_attention_heads_per_layer 配置。
- PR #10047 Llama4 non-uniform intermediate size（PR body 提及）：非均匀 per-layer 维度处理（Llama4 中间层大小）的既定模式，本 PR 的 get_hidden_dim hook 范式参考了该 PR。