PR #40411 完整报告

vllm-project/vllm

[Bugfix] Gemma4: fix multimodal embedder norm order to match HF reference

合并时间：2026-04-21 10:28

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40411

执行摘要

- 一句话：修正 Gemma4 多模态嵌入器中 LayerNorm 顺序
- 推荐动作：该 PR 值得精读，特别是对于想了解多模态模型中归一化位置对精度影响的研究者。类 docstring 的过时是一个微小残留问题，建议合入前或后续补上。

功能与动机

Gemma4MultimodalEmbedder 的实现与 HuggingFace transformers 参考实现存在差异：vLLM 中 RMSNorm 在 Linear 投影之后（post-projection），而 HF 参考实现在投影之前（pre-projection）。此差异导致视觉能力下降，修正后可提升 screen-spot 任务约 1.88pp 的 click accuracy。

实现拆解

1. 交换 __init__ 中的模块定义顺序和参数：将 embedding_post_projection_norm（RMSNorm，has_weight=False）改名为 embedding_pre_projection_norm，并将其 hidden_size 参数从 self.text_hidden_size 改为 embedding_dim（即投影前的维度）。原 embedding_projection（ReplicatedLinear）保持不变，但引用顺序调整。
2. 更新 forward 方法中的执行顺序：输入先经过 self.embedding_pre_projection_norm 归一化，再投影（self.embedding_projection），最后直接返回投影结果（不再额外归一化）。
3. 未触及 checkpoint 加载逻辑：因为 RMSNorm 的 has_weight=False，不包含可学习权重，所以无需修改权重映射。
4. 未添加新测试：改动仅涉及一个文件的内部逻辑，未增加测试。

关键文件：

- vllm/model_executor/models/gemma4_mm.py（模块 模型执行器；类别 source；类型 data-contract）：核心更改文件，修正 Gemma4MultimodalEmbedder 中 norm 的顺序

关键符号：未识别

关键源码片段

vllm/model_executor/models/gemma4_mm.py

核心更改文件，修正 Gemma4MultimodalEmbedder 中 norm 的顺序

```
class Gemma4MultimodalEmbedder(nn.Module):
```

```
    """Projects vision/audio soft tokens into LM embedding space.

    Architecture:
        inputs_embeds → embedding_pre_projection_norm → embedding_projection

    Unlike Gemma3n which has separate hard/soft embedding paths with
    per-path normalization and a learned embedding table, Gemma4 uses a
    simplified 2-layer design: pre-projection RMSNorm (without learnable scale)
    followed by a linear projection.  The checkpoint confirms this — only
    ``embedding_projection.weight`` exists; there is no embedding table
    or post-projection norm weights.
    """

    def __init__(
        self,
        multimodal_config: Gemma4VisionConfig | Gemma4AudioConfig,
        text_config: Gemma4TextConfig,
    ):
        super().__init__()

        self.eps = multimodal_config.rms_norm_eps
        self.text_hidden_size = text_config.hidden_size

        # Audio tower uses output_proj_dims (1536) rather than hidden_size
        # (1024); vision uses hidden_size (768) directly.
        embedding_dim = (
            getattr(multimodal_config, "output_proj_dims", None)
            or multimodal_config.hidden_size
        )

        # Apply RMSNorm BEFORE projection, matching HF reference.
        # The norm uses has_weight=False (no learnable scale),
        # so no checkpoint weight mapping changes are needed.
        self.embedding_pre_projection_norm = RMSNorm(
            embedding_dim,
            eps=self.eps,
            has_weight=False,
        )

        self.embedding_projection = ReplicatedLinear(
            embedding_dim,
            self.text_hidden_size,
            bias=False,
        )

    def forward(self, inputs_embeds: torch.Tensor) -> torch.Tensor:
        """Project soft tokens from a multimodal tower into LM space."""
        embs_normed = self.embedding_pre_projection_norm(inputs_embeds)
        embs_proj, _ = self.embedding_projection(embs_normed)
```

```
        return embs_proj
```

评论区精华

Code review 中 gemini-code-assist[bot] 指出类 docstring 中描述的架构 (inputs_embeds → embedding_projection → embedding_post_projection_norm) 与实际实现已不匹配，但作者未更新 docstring。该评论暂未解决，存在轻微文档滞后。

- 类 docstring 未更新 (documentation): 未在合并前解决；属于轻微文档滞后，不影响功能。

风险与影响

- 风险：低风险：
  - 仅为归一化顺序调整，不涉及权重映射或数据集加载变更。
  - 影响范围为 Gemma4MultimodalEmbedder 类，仅作用于视觉和音频软 token 嵌入，不影响语言模型主体。
  - 已通过 ScreenSpot-Pro 评测（1580 samples）验证效果提升，未出现退化迹象。
  - 影响：对用户：Gemma4 多模态模型（视觉 / 音频）的用户将获得约 1.88pp 的点击准确率提升。 对系统：无性能或内存影响（计算量相同）。 对团队：无部署或兼容性要求，可安全合入。
- 风险标记：模型正确性，核心路径变更，缺少测试覆盖

关联脉络

- 暂无明显关联 PR