PR #42081 完整报告

vllm-project/vllm

[Bug] Fix kimi dtype issue with `mm_projector_forward`

合并时间：2026-05-11 23:45

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42081

执行摘要

- 一句话：修复 Kimi K2.6 mm_projector 输入 dtype 不匹配崩溃
- 推荐动作：建议精读该 PR，了解多模态模型中自定义 forward 函数与 batch invariance 交互时可能的 dtype 问题。设计上，从 projector 的权重 dtype 推断预期输入 dtype 是合理做法，但可考虑更通用的契约（如所有涉及预处理的函数都显式转换）。

功能与动机

修复 Issue #41638：Kimi K2.6 模型在启用 batch invariance 模式时，因 mm_projector 的权重 dtype 为 Float 而输入为 BFloat16，导致 'expected scalar type Float but found BFloat16' 错误。PR body 提供了完整的复现命令和错误栈。

实现拆解

1. 修改文件：vllm/model_executor/models/kimi_k25_vit.py。
2. 获取目标 dtype：在 mm_projector_forward 函数中，通过 mm_projector.pre_norm.weight.dtype 获取 projector 期望的数据类型。
3. 类型检查与转换：检查输入 batched 张量的 dtype 是否与 projector_dtype 一致，如果不一致则调用 .to(projector_dtype) 进行转换。
4. 保持原有逻辑：后续的 projector 前向、reshape 和 split 逻辑不变。

关键文件：

- vllm/model_executor/models/kimi_k25_vit.py（模块 模型层；类别 source；类型 data-contract；符号 mm_projector_forward）：唯一修改的文件。在 mm_projector_forward 函数中新增 dtype 检查和转换逻辑，修复了 Kimi K2.6 的 batch invariance 崩溃。

关键符号：mm_projector_forward

关键源码片段

vllm/model_executor/models/kimi_k25_vit.py

唯一修改的文件。在 mm_projector_forward 函数中新增 dtype 检查和转换逻辑，修复了 Kimi K2.6 的 batch invariance 崩溃。

```
# vllm/model_executor/models/kimi_k25_vit.py
```

```
@torch.inference_mode()
def mm_projector_forward(mm_projector: torch.nn.Module, vt_output: list[torch.Tensor]):
    """Apply MM projector to vision tower outputs."""
    num_embedding_list = [x.shape[0] for x in vt_output]
    batched = torch.cat(vt_output, dim=0)
    # 获取 projector 期望的 dtype（通常为 Float），避免因 dtype 不匹配导致的崩溃
    projector_dtype = mm_projector.pre_norm.weight.dtype
    if batched.dtype != projector_dtype:
        batched = batched.to(projector_dtype)
    proj_out = mm_projector(batched)
    proj_out = proj_out.reshape(-1, proj_out.shape[-1])
    proj_out = torch.split(proj_out, num_embedding_list)
    return proj_out
```

评论区精华

无 review 讨论。gemini-code-assist[bot] 自动评论确认变更正确；sfeng33 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅在 mm_projector_forward 中添加了 dtype 检查与转换，不会改变原有行为。但未添加配套测试，可能遗漏回归覆盖。此外，如果 pre_norm 不存在或在其他模型上调用此函数，属性访问可能失败，但该函数目前仅用于 Kimi K2.6 系列。
- 影响：影响范围限于 Kimi K2.6 模型，具体为视觉编码器部分。修复后，启用 batch invariance 或 data parallel 的 Kimi K2.6 部署不再因 dtype 不匹配而崩溃，提升了该模型的推理稳定性。对系统其他部分无影响。
- 风险标记：代码变更未经测试

关联脉络

- PR #40408 [Perf] Batch invariance with Cutlass fp8 support, 28.9% E2E latency improvement: batch invariance 功能引入导致了 dtype 不匹配问题；本 PR 修复了该功能在 Kimi K2.6 上的 bug。