PR #30241 完整报告

sgl-project/sglang

[diffusion] Fix ragged-caption dynamic-batching accuracy bug in Ernie-Image

合并时间：2026-07-07 08:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/30241

执行摘要

- 一句话：修复 Ernie-Image 动态批处理变长标题精度错误
- 推荐动作：值得精读。PR 展示了如何系统性地修复动态批处理中变长输入的精度问题，包括正确的 mask 传递模式（从 postprocess 到 cond kwargs 再到 attention layer）。同时，作者通过逐步提交和 revert 管理风险的做法值得借鉴。重点关注 _prepare_encoder_hidden_states_mask 的设计（均匀长度返回 None）以及 build_varlen_mask_meta 在 DiT 中的使用。

功能与动机

PR #29742 修复了 Z-Image 中类似 bug，review 询问该 bug 是否 Z-Image 特有。分析发现 Ernie-Image 和 Wan T2V 存在相同漏洞：动态批处理合并不同长度请求后，DiT 未屏蔽填充 token 导致生成质量下降。本 PR 修复 Ernie-Image，并计划修复 Wan（后因回归恢复）。

实现拆解

1. 修改 tokenizer padding 配置：在 ernie_image.py 的 text_encoder_extra_args 中将 padding 从 False 改为 "longest"，使不同长度的请求可以合并为单个张量。
2. 重写 postprocess_text：ernie_image_postprocess_text 现在通过 _text_inputs.attention_mask 提取每个请求的真实（未填充）token 跨度，并调用 pad_text_embeddings_with_mask 返回 TextConditioningOutput，其中包含真实的 prompt_seq_lens 和 prompt_embeds_mask。
3. 新增掩码构建方法：在 ErnieImagePipelineConfig 中新增 _prepare_encoder_hidden_states_mask，当批处理中请求长度不一致时生成 [B, padded_len] 的 bool 掩码，长度一致时返回 None（零开销）。
4. 修改 _prepare_cond_kwargs：使用 require_text_seq_lens 获取真实长度，并调用 _prepare_encoder_hidden_states_mask 将掩码放入 cond_kwargs。
5. 修改 DiT 模型：在 ernie_image.py 的 ErnieImageSelfAttention.forward、ErnieImageBlock.forward 和 ErnieImageDiT.forward 中添加 attn_mask 和 attn_mask_meta 参数。DiT forward 中构建联合 [image, text] 掩码并调用 build_varlen_mask_meta，然后沿调用链传递到 USPAttention。
6. 新增单元测试：test_ernie_image_pipeline_config.py 包含两个测试类，分别验证 postprocess_text 正确提取真实长度和 _prepare_cond_kwargs 构建正确掩码（均匀长度

返回 None，变长返回边界掩码）。

关键文件：

- python/sglang/multimodal_gen/configs/pipeline_configs/ernie_image.py（模块 管道配置；类别 source；类型 core-logic；符号 ernie_image_postprocess_text, _prepare_cond_kwargs, _prepare_encoder_hidden_states_mask, ErnieImagePipelineConfig）：核心配置修改：tokenizer padding、postprocess_text 重构、新增掩码构建逻辑、修改 cond_kwargs 准备。
- python/sglang/multimodal_gen/test/unit/test_ernie_image_pipeline_config.py（模块 配置测试；类别 test；类型 test-coverage；符号 TestErnieImagePostprocessText, test_single_request_returns_full_length_conditioning, test_ragged_batch_preserves_real_lengths, TestErnieImagePrepareCondKwargs）：新增单元测试，覆盖 postprocess 的变长批处理和 cond_kwargs 的掩码逻辑。
- python/sglang/multimodal_gen/runtime/models/dits/ernie_image.py（模块 模型定义；类别 source；类型 data-contract；符号 ErnieImageSelfAttention.forward, ErnieImageBlock.forward, ErnieImageDiT.forward）：DiT 模型文件接受并传递注意力掩码到 self-attention 层，是数据契约变更的主要位置。

关键符号：ernie_image_postprocess_text, _prepare_encoder_hidden_states_mask, _prepare_cond_kwargs, ErnieImageSelfAttention.forward, ErnieImageBlock.forward, ErnieImageDiT.forward

关键源码片段

python/sglang/multimodal_gen/configs/pipeline_configs/ernie_image.py

核心配置修改：tokenizer padding、postprocess_text 重构、新增掩码构建逻辑、修改 cond_kwargs 准备。

```
def ernie_image_postprocess_text(outputs, _text_inputs, hidden_layer_index=-2):
    """Return Ernie-Image text embeddings, re-padded from real token spans.
    Batched requests can have different real caption lengths after
    tokenization; extract each request's real (unpadded) span via the
    tokenizer's attention mask and re-pad, so TextConditioningOutput carries
    the true per-request lengths instead of the tokenizer's padded length.
    """
    hidden_states = outputs.hidden_states[hidden_layer_index]
    prompt_mask = _text_inputs.attention_mask.to(hidden_states.device).bool()
    split_hidden_states = [
        hidden_states[idx][prompt_mask[idx]] for idx in range(hidden_states.shape[0])
    ]
    # pad_text_embeddings_with_mask 返回 TextConditioningOutput，包含
    # prompt_embeds（重新填充后的 [B, P, D]）、prompt_seq_lens（真实长度列表）
    # 和 prompt_embeds_mask (bool 掩码)
    return pad_text_embeddings_with_mask(split_hidden_states)


def _prepare_encoder_hidden_states_mask(
```

```
    self,
    batch,
    txt_seq_lens: list[int],
    text_seq_len: int,
    device,
):
    """Return a `[batch, text_seq_len]` bool mask over real (non-padded) text tokens.
    Dynamic batches can merge requests whose captions have different real
    lengths after tokenization; the DiT still sees one padded
    `encoder_hidden_states` tensor of shape `[batch, text_seq_len, dim]`, so
    we need a mask to keep attention off the padding. Returns None when every
    request already fills the full padded length (no mask needed, zero overhead).
    """
    if all(seq_len == text_seq_len for seq_len in txt_seq_lens):
        # 均匀长度：掩码为 None，注意力层不执行任何遮盖
        return None

    positions = torch.arange(text_seq_len, device=device)
    seq_lens = torch.tensor(txt_seq_lens, device=device, dtype=torch.long)
    # positions: [0, 1, ..., text_seq_len-1]
    # 对每个位置 i，若 i < 该请求的 real seq_len，则标记为 True
    return positions.unsqueeze(0) < seq_lens.unsqueeze(1)
```

评论区精华

代码审查机器人 gemini-code-assist[bot] 提出了两个关键问题：

- WanI2VCrossAttention context_lens 形状不匹配（严重）：context_lens 形状为 [B, 512] 代表文本部分，但代码中错误地切片 [:, 257:] 导致形状不匹配。建议直接使用整个掩码。
- t5_postprocess_text 设备不匹配风险（中等）：seq_lens 可能不在 positions 所在设备，建议显式 .to(positions.device)。 作者后来恢复 Wan 部分变更，这两个问题随 Wan 代码撤回而消失，Ernie-Image 部分保持正确。
- WanI2VCrossAttention context_lens 形状不匹配 (correctness): 作者随后恢复 Wan 部分变更，因此该问题无需修复（Wan 不再改动）。
- t5_postprocess_text 设备不匹配风险 (correctness): 此问题随 Wan 恢复而消失，未在最终代码中体现。

风险与影响

- 风险：
  - 核心路径变更风险：DiT forward 签名新增两个参数，需确认无其他直接调用 ErnieImageDiT.forward() 的地方（通过 CPP 或外部脚本）。
  - 缺少 GPU 端到端验证：作者明确标注未运行 GPU 测试，存在潜行回归风险。CI 的 pre-existing 依赖冲突也阻止了单元测试在本环境中执行。

    - Wan 部分恢复暴露风险：Wan 的掩码导致严重视觉回归，表明类似逻辑可能引入意想不到的精度影响。虽然 Ernie-Image 已独立测试，但缺乏 GPU 验证仍需谨慎。
    - 单元测试覆盖有限：仅覆盖 CPU 侧的掩码构建逻辑，未测试 DiT forward 的实际注意力行为。
- 影响：
    - 用户影响：Ernie-Image 模型动态批处理生成质量显著提升，消除了变长标题批处理时生成的静默错误。单请求或均匀长度批处理不受影响（掩码为 None，零额外开销）。
    - 系统影响：对均匀长度请求无开销，对变长批处理引入少量掩码计算和 build_varlen_mask_meta 开销，但注意力计算本身节省了填充 token 的计算（通过 sparse attention）。
    - 团队影响：展示了可复用的变长掩码模式（require_text_seq_lens / build_varlen_mask_meta），为后续修复 SD3 等模型提供了参考。
    - 风险标记：核心路径变更，缺少 GPU 端到端验证，Wan 部分恢复，测试覆盖有限

关联脉络

- PR #29742 Fix Z-Image ragged-caption dynamic-batching accuracy bug: 本 PR 是 #29742 的后续，修复了同类漏洞在 Ernie-Image 中的表现，并使用了相同的修复模式（TextConditioningOutput、require_text_seq_lens）。