PR #35829 完整报告

sgl-project/sglang

[diffusion] feat: support LongCat-Image-Edit and LongCat-Image-Edit-Turbo

合并时间：2026-08-25 12:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35829

执行摘要

- 一句话：新增 LongCat-Image-Edit 及 Turbo 图像编辑管线
- 推荐动作：值得精读，尤其是 honor_cache_free_padding_mask 的作用域设计、expand_conditioning_to_sample_batch 的多输出处理，以及 _build_expanded_prefix_ids 的占位符展开逻辑。建议关注共享编码器开关的后续加固和 GPU 端到端测试的补充。

功能与动机

PR body 明确说明：SGLang diffusion currently supports LongCat-Image for text-to-image only. This PR adds image-editing (I2I) support for meituan-longcat/LongCat-Image-Edit and the distilled meituan-longcat/LongCat-Image-Edit-Turbo, mirroring the diffusers LongCatImageEditPipeline reference implementation.

实现拆解

1. 配置层：在 python/sglang/multimodal_gen/configs/pipeline_configs/longcat_image.py 新增 LongCatImageEditPipelineConfig，继承 T2I 配置并覆盖关键 hooks。_calculate_edit_dimensions 按 diffusers 公式计算约 1MP 面积并向上取整到 /16；slice_noise_pred 去噪后丢弃参考图 token 的预测；maybe_prepare_latent_ids 返回 None（位置 ID 在每步动态生成）；新增 expand_conditioning_to_sample_batch 处理多输出条件展开。
2. 编码阶段：新增 LongCatImageEditTextEncodingStage（runtime/pipelines_core/stages/model_specific_stages/longcat_image_edit.py），将参考图缩小 2 倍送入 Qwen2.5-VL 视觉塔，把编辑 system prompt 中的 <|image_pad|> 展开为实际图像 token 数，对正文做引号感知的 512 token 截断，输出隐藏状态从 <|vision_start|> 切片到 512-token 正文；CFG 开启时负向 prompt 也基于同一参考图编码。
3. 管线组装：在 runtime/pipelines/longcat_image.py 新增 LongCatImageEditPipeline，串联 InputValidationStage（加载并缩放条件图）、自定义文本编码 stage、ImageVAEEncodingStage（参考图 argmax VAE 编码）、标准 latent 准备、时间步准备、去噪循环和 VAE 解码；参考图 latent 沿序列维拼接到噪声 latent 之后，每步预测后由 slice_noise_pred 切掉参考 token。

4. 注册与默认参数：在 registry.py 注册两个模型 ID；在 configs/sample/longcat_image.py 定义 LongCatImageEditSamplingParams（50 步、CFG 4.5、空负向提示）和 LongCatImageEditTurboSamplingParams（8 步、CFG off），Turbo 检测器先注册以确保优先匹配。
5. 共享编码器适配：runtime/models/encoders/qwen2_5vl.py 新增 _build_causal_padding_mask，并通过 honor_cache_free_padding_mask 配置开关控制 cache-free 路径是否传递 attention mask；text_encoder_loader.py 仅对 LongCat 相关 config 设置该开关，避免 Qwen-Image 等其它管线从 mask-free 快速路径切到 SDPA。
6. 测试与文档：新增 test/unit/test_longcat_image_edit_config.py（CPU-only）覆盖尺寸公式、切片、位置 ID、条件展开等；更新 test_qwen2_5vl_generation.py 验证开关两种状态；更新 docs/docs/sglang-diffusion/compatibility_matrix.mdx。

关键文件：

- python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/longcat_image_edit.py（模块 编码阶段；类别 source；类型 data-contract；符号 LongCatImageEditTextEncodingStage, _get_suffix_ids, _build_expanded_prefix_ids, _encode）：新增的联合 VL prompt 编码 stage，是 I2I 管线核心逻辑，负责将编辑指令与参考图编码为 DiT 条件向量。
- python/sglang/multimodal_gen/configs/pipeline_configs/longcat_image.py（模块 管线配置；类别 source；类型 dependency-wiring；符号 LongCatImageEditPipelineConfig, _calculate_edit_dimensions, expand_conditioning_to_sample_batch, slice_noise_pred）：新增 I2I 配置类与尺寸计算、条件展开等 hooks，是管线行为的重要控制点。
- python/sglang/multimodal_gen/test/unit/test_longcat_image_edit_config.py（模块 单元测试；类别 test；类型 test-coverage；符号 test_edit_dimensions_match_diffusers_formula, test_slice_noise_pred_drops_reference_tokens, test_edit_img_ids_modalities_and_offset, test_expand_conditioning_repeats_embeds_for_num_outputs）：新增 CPU-only 单元测试，覆盖配置 hooks 的正确性，是防止回归的关键保障。
- python/sglang/multimodal_gen/runtime/pipelines/longcat_image.py（模块 管线组装；类别 source；类型 core-logic；符号 LongCatImageEditPipeline, create_pipeline_stages）：新增 LongCatImageEditPipeline 的组装逻辑，定义 I2I 管线各 stage 顺序。
- python/sglang/multimodal_gen/runtime/models/encoders/qwen2_5vl.py（模块 文本编码器；类别 source；类型 data-contract；符号 _build_causal_padding_mask, honor_cache_free_padding_mask）：共享 Qwen2.5-VL 编码器新增 cache-free padding mask 支持，并用开关限定到 LongCat，是影响面控制的关键。
- python/sglang/multimodal_gen/configs/sample/longcat_image.py（模块 采样参数；类别 source；类型 core-logic；符号 LongCatImageEditSamplingParams, LongCatImageEditTurboSamplingParams）：定义 Edit 和 Turbo 的采样默认值，影响用户默认行为。
- python/sglang/multimodal_gen/registry.py（模块 模型注册；类别 source；类型 core-logic）：注册新模型到模型注册表，是模型可被识别和调用的入口。

- docs/docs/sglang-diffusion/compatibility_matrix.mdx（模块 文档；类别 other；类型 documentation）：更新兼容性矩阵文档，便于用户查看支持的模型。

关键符号：LongCatImageEditPipeline, LongCatImageEditPipelineConfig, LongCatImageEditTextEncodingStage, _build_expanded_prefix_ids, _encode, expand_conditioning_to_sample_batch, _calculate_edit_dimensions, _build_causal_padding_mask, slice_noise_pred, maybe_prepare_latent_ids, postprocess_image_latent, _edit_img_ids

关键源码片段

python/sglang/multimodal_gen/runtime/pipelines_core/stages/model_specific_stages/longcat_image_edit.py

新增的联合 VL prompt 编码 stage，是 I2I 管线核心逻辑，负责将编辑指令与参考图编码为 DiT 条件向量。

```
def _build_expanded_prefix_ids(self, image_grid_thw: torch.Tensor) -> tuple[list[int], int]:
    """将编辑系统前缀中的 <|image_pad|> 占位符展开为真实图像 token 数。

    返回前缀 token 列表和 <|vision_start|> 的位置索引 prefix_len；
    后续编码输出会从 prefix_len 处切片，保证 DiT 条件包含 VL 图像 token。
    """
    # merge_size 用于把视觉特征合并为图像 token 数量
    merge_length = self.text_processor.image_processor.merge_size**2
    num_image_tokens = int(image_grid_thw.prod().item()) // merge_length
    text = PROMPT_TEMPLATE_ENCODE_PREFIX
    # 先用唯一占位符逐个替换，再统一换回 IMAGE_TOKEN，避免多次 replace 互相影响
    while IMAGE_TOKEN in text:
        text = text.replace(IMAGE_TOKEN, "<|placeholder|>" * num_image_tokens, 1)
    text = text.replace("<|placeholder|>", IMAGE_TOKEN)

    prefix_ids = self.tokenizer(text, add_special_tokens=False)["input_ids"]
    vision_start_id = self.tokenizer.convert_tokens_to_ids("<|vision_start|>")
    prefix_len = prefix_ids.index(vision_start_id)
    return prefix_ids, prefix_len
```

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

新增 I2I 配置类与尺寸计算、条件展开等 hooks，是管线行为的重要控制点。

```
def expand_conditioning_to_sample_batch(self, batch):
    """当 num_outputs_per_prompt > 1 时，把 per-prompt 文本条件展开到 batch 维度。

    噪声/参考图 latent 以 prompts * num_outputs 构建，文本编码仍按 prompt
    粒度，所以需要重复 embeds、masks、seq_lens；n=1 时是 no-op。
    """
    from sglang.multimodal_gen.runtime.utils.condition_expansion import (
        PromptToSampleBatchExpander,
    )
```

```
    expander = PromptToSampleBatchExpander.from_batch(batch)
    if expander is None:
        return batch
    for field_name in (
        "prompt_embeds",
        "negative_prompt_embeds",
        "prompt_embeds_mask",
        "negative_prompt_embeds_mask",
        "prompt_seq_lens",
        "negative_prompt_seq_lens",
    ):
        expander.expand_field(batch, field_name)
    return batch


def _calculate_edit_dimensions(target_area, ratio):
    """计算 LongCat-Image-Edit 输出尺寸：拟合目标面积后向上取整到 16 的倍数。

    注意与 sglang.multimodal_gen.utils.calculate_dimensions 不同，后者按 /32
    对齐，而 diffusers 参考实现按 /16 对齐（ceil），因此这里单独实现。
    """
    width = math.sqrt(target_area * ratio)
    height = width / ratio

    width = width if width % 16 == 0 else (width // 16 + 1) * 16
    height = height if height % 16 == 0 else (height // 16 + 1) * 16

    return int(width), int(height)
```

评论区精华

1. 多输出 batch 不匹配：BBuf 指出 num_outputs_per_prompt > 1 时 prompt_embeds 保持 batch 1，与 latents 的 batch 不一致，ITerydh 随后增加 expand_conditioning_to_sample_batch 修复并补充 n=2 测试。
2. T2I 行为变化争议：BBuf 指出 Qwen2.5-VL mask 改动会使 LongCat T2I 行为变化，与 PR 描述“unchanged”不符；ITerydh 承认这是向 diffusers parity 的修复，并补充了 parity 验证。
3. 作用域限制：BBuf 要求将 mask 改动限定到 LongCat，避免切换 LocalAttention 到 SDPA 影响 Qwen-Image；ITerydh 通过 honor_cache_free_padding_mask 开关实现限定，并验证 LongCat 输出字节一致。

- num_outputs_per_prompt > 1 时条件展开缺失 (correctness): ITerydh 添加了 expand_conditioning_to_sample_batch 覆盖，镜像 QwenImagePipelineConfig，并增加 n=2 测试，端到端验证通过。
- Qwen2.5-VL cache-free mask 改变 LongCat T2I 行为 (correctness): ITerydh 承认这是向 diffusers parity 的修复，并做了 parity 验证；随后在 BBuf 要求下将该行为限定到 LongCat（honor_cache_free_padding_mask）。

- 共享编码器改动的作用域限制 (design): ITerydh 通过 TextEncoderLoader 仅对 LongCat 设置 honor_cache_free_padding_mask 开关，其他 pipeline 保持原路径，并验证 LongCat 输出字节一致。

风险与影响

- 风险：
  1. 共享编码器风险：qwen2_5vl.py 的改动影响所有使用该编码器的管线，虽已用开关隔离，但未来若其他模型误设 flag 可能静默切换 attention 实现，建议在 loader 层加固。
  2. T2I 行为微调：LongCat T2I 原本保留 padding token 参与注意力，现在被 mask 掉，输出会变化；虽验证为向 diffusers 收敛的正确修复，但用户可能需要更新期望值。
  3. 测试覆盖不足：新增测试均为 CPU 级 config hook，缺少 GPU 端到端自动化，后续改动 DiT 位置 ID 或 VAE 处理逻辑回归风险较高。
  4. 多输出字段覆盖：expand_conditioning_to_sample_batch 只覆盖列出的 6 个字段，未来新增条件字段需同步扩展。 - 影响：对用户而言，新增了两个模型的 I2I 支持，扩展了 diffusion 能力；对系统而言，共享 Qwen2.5-VL 编码器新增一个默认关闭的开关，其他管线无行为影响，但 LongCat T2I 输出会向 diffusers 收敛（微小变化）；对团队而言，为后续编辑类 diffusion pipeline 提供了可复用的模式，如条件展开、VAE 编码 stage 复用。 - 风险标记：共享编码器增加开关，LongCat T2I 行为微调，GPU E2E 测试缺失，diffusers 对齐依赖

关联脉络

- PR #36066 [diffusion] feat: dispatch fp8 companions in mixed nvfp4 checkpoints: 同属 diffusion runtime 的功能扩展，涉及共享的量化 / 加载逻辑，与本 PR 在 multimodal_gen 子模块有重叠。
- PR #36046 [Diffusion] Load Comfy NVFP4-AWQ text encoders: 同样修改了 text encoder 加载与量化路径，可能与 Qwen2.5-VL 编码器交互。
- PR #33859 [diffusion] fix: crop GLM-Image output to requested size: 同为 diffusion 特定 pipeline 的修复 / 支持，说明该模块持续演进，可对比学习 pipeline 配置 hooks 的扩展方式。