PR #41759 完整报告

vllm-project/vllm

[MM][Perf][CG] Support ViT full CUDA graph for InternVL

合并时间：2026-06-04 10:24

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/41759

执行摘要

- 一句话：为 InternVL 系列添加 ViT CUDA 图支持
- 推荐动作：值得精读。PR 展示了如何为 ViT 编码器集成 CUDA 图，包括协议方法实现、测试和文档配套。特别关注接口适配 #41234 的过程，以及如何解决 MIG 环境兼容性问题。

功能与动机

ViT 编码器前向传播包含大量小 CUDA 内核，每次推理均重新启动，产生显著开销（关联 Issue #38175 跟踪此问题）。本 PR 将 CUDA 图能力扩展到 InternVL 系列，减少内核启动开销，提升多模态模型推理性能。

实现拆解

1. 在 vllm/model_executor/models/internvl.py 中，让 InternVLChatModel 继承 SupportsEncoderCudaGraph 协议，并实现：get_encoder_cudagraph_config（定义模态和 buffer 键）、get_input_modality（从 mm_kwargs 判断模态）、get_encoder_cudagraph_budget_range（预算范围）、get_encoder_cudagraph_item_specs（项目输入 / 输出规格）、select_encoder_cudagraph_items（选择项目）和 prepare_encoder_cudagraph_capture_inputs（准备捕获输入）。同时移除了过时的 supports_encoder_cudagraph 类属性和 get_max_frames_per_video 重写。
2. 在 tests/models/multimodal/generation/test_vit_cudagraph.py 中，添加 internvl_chat_template 辅助函数和 InternVL3-1B 的 VitCudagraphTestConfig，覆盖图像和视频模态，使用 pytest.mark.core_model 标记。
3. 在 examples/generate/multimodal/vision_language_offline.py 的 MODELS_SUPPORT_VIT_CUDA_GRAPH 列表中增加 internvl_chat 条目。
4. 在 docs/design/cuda_graphs_multimodal.md 的模型表格中增加 InternVLChatModel 行，注明对图像和视频均支持。

关键文件：

- vllm/model_executor/models/internvl.py（模块 模型执行器；类别 source；类型 data-contract；符号 InternVLChatModel, get_encoder_cudagraph_config, get_input_modality, get_encoder_cudagraph_budget_range）：核心实现。添加 SupportsEncoderCudaGraph 继承，并实现协议方法：get_encoder_cudagraph_config,

get_input_modality, get_encoder_cudagraph_budget_range,get_encoder_cudagraph_item_specs, select_encoder_cudagraph_items,prepare_encoder_cudagraph_capture_inputs。同时移除不必要的类属性和过时方法。

- tests/models/multimodal/generation/test_vit_cudagraph.py（模块 CUDA 图；类别 test；类型 test-coverage；符号 internvl_chat_template）：添加 InternVL 测试用例，包括 internvl_chat_template 函数和 InternVL3-1B 的 VitCudagraphTestConfig，覆盖图像和视频模态。
- examples/generate/multimodal/vision_language_offline.py（模块 示例；类别 source；类型 core-logic）：在 MODELS_SUPPORT_VIT_CUDA_GRAPH 列表中注册 internvl_chat，使示例脚本能使用 CUDA 图。
- docs/design/cuda_graphs_multimodal.md（模块 文档；类别 docs；类型 documentation）：更新文档，在支持 CUDA 图的模型表格中添加 InternVLChatModel 行，注明图像和视频均支持。

关键符号：get_encoder_cudagraph_config, get_input_modality, get_encoder_cudagraph_budget_range, _get_internvl_patches_list, get_encoder_cudagraph_item_specs, select_encoder_cudagraph_items, prepare_encoder_cudagraph_capture_inputs, internvl_chat_template

关键源码片段

vllm/model_executor/models/internvl.py

核心实现。添加 SupportsEncoderCudaGraph 继承，并实现协议方法：get_encoder_cudagraph_config, get_input_modality, get_encoder_cudagraph_budget_range, get_encoder_cudagraph_item_specs, select_encoder_cudagraph_items, prepare_encoder_cudagraph_capture_inputs。同时移除不必要的类属性和过时方法。

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

class InternVLChatModel(
    nn.Module,
    SupportsMultiModal,
    SupportsPP,
    SupportsLoRA,
    SupportsEncoderCudaGraph, # 新加入的协议，表明支持 ViT CUDA 图
):
    supports_encoder_tp_data = True

    # -- SupportsEncoderCudaGraph protocol methods --

    def get_encoder_cudagraph_config(self):
        from vllm.v1.worker.encoder_cudagraph_defs import EncoderCudaGraphConfig
        # InternVision 使用标准 ViT attention（无 rotary embedding，
        # 无可变长度序列元数据），所以唯一需要的 graph buffer 是
        # pixel_values_flat 本身。
```

```
        return EncoderCudaGraphConfig(
            modalities=["image", "video"],
            buffer_keys=["pixel_values_flat"],
            out_hidden_size=self.config.text_config.hidden_size,
        )

    def get_input_modality(
        self,
        mm_kwargs: dict[str, Any],
    ) -> str:
        if "pixel_values_flat" in mm_kwargs:
            return "image"
        return "video"

    def get_encoder_cudagraph_budget_range(
        self,
        vllm_config: "VllmConfig",
    ) -> tuple[int, int]:
        # 最小预算：1 个 tile 对应的 num_image_token 个输出 token
        min_budget = self.num_image_token
        max_budget = min(
            vllm_config.scheduler_config.max_num_batched_tokens,
            vllm_config.model_config.max_model_len,
        )
        return (min_budget, max_budget)

    def get_encoder_cudagraph_item_specs(
        self,
        mm_kwargs: dict[str, Any],
    ):
        from vllm.v1.worker.encoder_cudagraph_defs import EncoderItemSpec
        return [
            EncoderItemSpec(
                input_size=n,
                output_tokens=n * self.num_image_token,
            )
            for n in self._get_internvl_patches_list(mm_kwargs)
        ]
```

tests/models/multimodal/generation/test_vit_cudagraph.py

添加 InternVL 测试用例，包括 internvl_chat_template 函数和 InternVL3-1B 的 VitCudagraphTestConfig，覆盖图像和视频模态。

```
# tests/models/multimodal/generation/test_vit_cudagraph.py

def internvl_chat_template(content: str) -> str:
    # InternVL 使用的 chat template 与 Qwen 相同
    return f"<|im_start|>user\n{content}<|im_end|>\n<|im_start|>assistant\n"
```

```
MODEL_CONFIGS: dict[str, VitCudagraphTestConfig] = {
    "internvl": VitCudagraphTestConfig(
        model="OpenGVLab/InternVL3-1B",
        num_video_frames=8,
        image_prompt=internvl_chat_template("<image>\nWhat is in this image?"),
        video_prompt=internvl_chat_template(
            "<video>\nDescribe this video in one sentence."
        ),
        needs_video_metadata=False,
        vllm_runner_kwargs={"trust_remote_code": True},
        marks=[pytest.mark.core_model],
    ),
    # 其他模型配置保持不变 ...
}
```

评论区精华

核心讨论围绕接口适配和简化：@Isotr0py 指出需要适配 #41234 引入的新接口（get_encoder_cudagraph_item_specs 等），作者及时更新。@shen-shanshan 询问 InternVL3.5 支持情况，作者确认并更新文档；建议使用更小的模型（InternVL3-1B）以减少测试资源，作者采纳；指出 supports_encoder_cudagraph 类属性和 get_max_frames_per_video 重写不必要（默认值已满足），作者移除。此外，作者在评论中报告了 MIG 环境下的 CUDA 内存分配器 NVML assert 问题，通过减少视频帧数规避。

- 需要适配 #41234 新接口 (design): 作者在后续提交中更新了接口，替换了三个遗留方法。
- InternVL3.5 支持确认 (question): 作者确认 InternVL3.5 也使用 InternVLChatModel 架构，因此支持，并更新了文档表格。
- 测试使用更小模型 (testing): 作者采纳，将测试模型从 2B 改为 1B，并相应调整视频帧数。
- 移除不必要的类属性 (style): 作者移除了这些冗余代码。

风险与影响

- 风险：主要风险包括： 1) CUDA 图捕获在部分 GPU 环境（如 MIG）下可能触发 NVML assert（已在 H200 MIG 上复现），虽然通过调小测试规模规避，但生产环境中若使用大尺寸输入或高并发可能复现； 2) 新引入的 SupportsEncoderCudaGraph 协议接口在后续版本中可能继续演进，后续维护需要同步更新； 3) InternVL 的 select_encoder_cudagraph_items 中区分了 pixel_values_flat 和 pixel_values_flat_video 键，若未来 InternVL 的输入键发生变化，需要相应调整。
  - 影响：对用户：使用 InternVL3、InternVL2.5、InternVL2 的用户将自动获得 ViT 编码器 CUDA 图加速，无需额外配置，TTFT 降低 2%-17%。对系统：CUDA 图捕获会占用少量显存（根据 budget 配置），但可提升 GPU 利用率。对团队：此 PR 展示了为多模态模型添加 CUDA 图支持的标准模式（实现 SupportsEncoderCudaGraph 协议），后续其他模型（如 DeepSeek VL、Qwen3.6 等）可参照实现。
  - 风险标记：MIG 环境 NVML assert, CUDA 图捕获兼容性，协议接口演进依赖

关联脉络

- PR #38061 [MM][Perf][CG] Support ViT full CUDA graph for Qwen3-VL: 此 PR 的前驱，为 Qwen3-VL 实现 ViT CUDA 图，本 PR 采用相同模式。
- PR #41234 [Refactor] Introduce EncoderItemSpec and update interfaces: 引入新的 EncoderItemSpec 接口，本 PR 需适配此变更。
- PR #38175 [RFC]: Support ViT Full CUDA Graph (Tracker): 此 PR 是跟踪 issue 中列出的里程碑之一。
- PR #42288 [Refactor] Remove input_key_by_modality: 本 PR 最终同步了该变更，移除了 input_key_by_modality。