# PR #42151 完整报告

vllm-project/vllm

[MM][Perf][CG] Support ViT full CUDA graph for Qwen3.5

合并时间：2026-05-13 16:00

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42151

## 执行摘要

- 一句话：为 Qwen3.5 启用 ViT 全 CUDA 图
- 推荐动作：建议合并。该 PR 遵循了已建立的 ViT CUDA 图扩展模式，代码清晰，测试覆盖完善，审阅人批准。值得关注的设计决策：通过复用 Qwen3-VL 的视觉变换器，展示了 vLLM 中多模态模型 CUDA 图支持的模块化扩展方法。

## 功能与动机

根据 PR 标题和说明，目的是支持 Qwen3.5 的 ViT 全 CUDA 图（ViT full CUDA graph），复用 Qwen3-VL 的 ViT 实现，提升多模态场景下的视觉编码推理性能。

## 实现拆解

1. 保存 model_config 引用：在 vllm/model_executor/models/qwen3_5.py 中，为 Qwen3_5ForConditionalGeneration 和 Qwen3_5MoeForConditionalGeneration 的 __init__ 方法添加 self.model_config = vllm_config.model_config，使 CUDA 图框架能访问模型配置（如 encoder_cudagraph_max_vision_items_per_batch）。
2. 新增模型示例入口：在 examples/generate/multimodal/vision_language_offline.py 中定义 run_qwen3_5（使用 Qwen/Qwen3.5-4B）和 run_qwen3_5_moe（使用 Qwen/Qwen3.5-35B-A3B）两个函数，并添加到 MODELS_SUPPORT_VIT_CUDA_GRAPH 和 MODELS_NEED_VIDEO_METADATA 列表中。
3. 扩展 CUDA 图测试覆盖：在 tests/models/multimodal/generation/test_vit_cudagraph.py 中调整现有模型顺序，新增 qwen3_5 测试配置（使用 Qwen/Qwen3.5-0.8B），验证 image 和 video 模态下的 CUDA 图功能。
4. 更新文档：在 docs/design/cuda_graphs_multimodal.md 的支持模型表格中增加 Qwen3_5ForConditionalGeneration 行，明确图像和视频 CUDA 图支持。

关键文件：

- examples/generate/multimodal/vision_language_offline.py（模块 示例脚本；类别 source；类型 core-logic；符号 run_qwen3_5, run_qwen3_5_moe）：新增 Qwen3.5 Dense/MoE 模型示例入口，注册支持列表
- vllm/model_executor/models/qwen3_5.py（模块 模型执行器；类别 source；类型 data-contract）：在模型初始化中保存 model_config，是启用 CUDA 图的关键前提

- tests/models/multimodal/generation/test_vit_cudagraph.py（模块 测试；类别 test；类型 test-coverage）：添加 Qwen3.5 CUDA 图测试配置，确保功能验证
- docs/design/cuda_graphs_multimodal.md（模块 文档；类别 docs；类型 documentation）：更新支持模型表格，同步文档

关键符号：run_qwen3_5, run_qwen3_5_moe, Qwen3_5ForConditionalGeneration.init, Qwen3_5MoeForConditionalGeneration.init

## 关键源码片段

### examples/generate/multimodal/vision_language_offline.py

新增 Qwen3.5 Dense/MoE 模型示例入口，注册支持列表

```
def run_qwen3_5(questions: list[str], modality: str) -> ModelRequestData:
    # 设置模型名称与引擎参数，复用 Qwen3-VL 的处理器逻辑
    model_name = "Qwen/Qwen3.5-4B"
    mm_limit = {"image": 1, "video": 1} if modality == "image+video" else {modality: 1}
    engine_args = EngineArgs(
        model=model_name,
        max_model_len=4096,
        max_num_seqs=5,
        mm_processor_kwargs={
            "min_pixels": 28 * 28,
            "max_pixels": 1280 * 28 * 28,
            "fps": 1,
        },
        limit_mm_per_prompt=mm_limit,
    )
    # 构建占位符与 prompt，格式与 Qwen3-VL 一致
    image_placeholder = "<|vision_start|><|image_pad|><|vision_end|>"
    video_placeholder = "<|vision_start|><|video_pad|><|vision_end|>"
    if modality == "image":
        placeholder = image_placeholder
    elif modality == "video":
        placeholder = video_placeholder
    elif modality == "image+video":
        placeholder = image_placeholder + video_placeholder
    prompts = [
        "<|im_start|>system\nYou are a helpful assistant.<|im_end|>\n"
        f"<|im_start|>user\n{placeholder}{question}<|im_end|>\n"
        "<|im_start|>assistant\n"
        for question in questions
    ]
    return ModelRequestData(engine_args=engine_args, prompts=prompts)
```

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

添加 Qwen3.5 CUDA 图测试配置，确保功能验证

```
# 在 MODEL_CONFIGS 中新增 Qwen3.5 条目，使用 0.8B 小模型进行快速验证
```

```
MODEL_CONFIGS: dict[str, VitCudagraphTestConfig] = {
    "qwen2_5_vl": VitCudagraphTestConfig(
        model="Qwen/Qwen2.5-VL-3B-Instruct",
        image_prompt=qwen_vl_chat_template(
            "<|vision_start|><|image_pad|><|vision_end|>What is in this image?"
        ),
        video_prompt=qwen_vl_chat_template(
            "<|vision_start|><|video_pad|><|vision_end|>"
            "Describe this video in one sentence."
        ),
        needs_video_metadata=False,
        marks=[pytest.mark.core_model],
    ),
    "qwen3_vl": ... # 保持原有
    "qwen3_5": VitCudagraphTestConfig(
        model="Qwen/Qwen3.5-0.8B",
        image_prompt=qwen_vl_chat_template(
            "<|vision_start|><|image_pad|><|vision_end|>What is in this image?"
        ),
        video_prompt=qwen_vl_chat_template(
            "<|vision_start|><|video_pad|><|vision_end|>"
            "Describe this video in one sentence."
        ),
        needs_video_metadata=True, # Qwen3.5 需要视频元数据
        marks=[pytest.mark.core_model],
    ),
}
```

# 评论区精华

本 PR 没有实质性的 review 讨论。GitHub 自动审查机器人（claude、gemini-code-assist）未提出具体问题，审核人 Isotr0py 直接批准，说明变更清晰、风险低。

- 暂无高价值评论线程

# 风险与影响

- 风险：风险较低。改动量小（共 112 行增加，5 行删除），且严格遵循已有 Qwen3-VL 的 CUDA 图模式。主要风险点包括：(1) Qwen3.5 的视觉变换器与 Qwen3-VL 完全兼容，但未来 upstream 更新可能产生差异，需同步适配；(2) 测试依赖远程 HuggingFace 模型权重，在网络不稳定时可能失败；(3) model_config 存储增加了一个引用字段，无性能影响。
- 影响：影响范围：仅限于 Qwen3.5 模型用户。启用 --enable-vit-cuda-graph 后，视觉编码器推理性能显著提升（在 Blackwell 等 GPU 上已验证）。向后兼容：无需额外迁移工作，现有配置不受影响。
- 风险标记：改动量小，测试依赖远程模型权重

# 关联脉络

- 暂无明显关联 PR