PR #40780 完整报告

vllm-project/vllm

[CI/Build] Add e2e test for ViT CUDA graph

合并时间：2026-04-24 18:12

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40780

执行摘要

本 PR 为 ViT CUDA graph 添加了端到端测试，覆盖图片和视频两种模态，并集成到 CI 中。当前仅覆盖 Qwen3-VL 模型，但测试框架设计便于扩展。

功能与动机

ViT CUDA graph 是 vLLM 多模态推理的重要优化，但此前缺乏专门的端到端测试。本 PR 通过新增测试文件 test_vit_cudagraph.py，确保 CUDA graph 在 ViT 编码器上的正确工作，防止回归。

实现拆解

1. 测试配置类：定义 VitCudagraphTestConfig 数据类，集中管理模型、模态、提示、编译配置等参数。
2. 辅助函数：params_with_marks 将配置字典转化为 pytest 参数；qwen_vl_chat_template 生成 Qwen3-VL 聊天模板；get_compilation_config 返回启用 CUDA graph 的配置字典。
3. 测试函数：test_vit_cudagraph_image 和 test_vit_cudagraph_video 使用 vllm_runner 启动推理，验证输出非空且为字符串。
4. CI 集成：在 models_multimodal.yaml 的 qwen3+gemma 步骤中添加 pytest 命令。

关键代码片段如下：

tests/models/multimodal/generation/test_vit_cudagraph.py

新增测试文件，包含所有 ViT CUDA graph 端到端测试逻辑，是 PR 的核心变更。

关键源码片段

tests/models/multimodal/generation/test_vit_cudagraph.py

新增测试文件，包含所有 ViT CUDA graph 端到端测试逻辑，是 PR 的核心变更。

```
@dataclass
class VitCudagraphTestConfig:
    model: str
    modalities: list[str] = field(default_factory=lambda: ["image", "video"])
    image_prompt: str | None = None
    video_prompt: str | None = None
    dtype: str = "bfloat16"
```

```
    max_model_len: int = 4096
    max_tokens: int = 64
    max_num_seqs: int = 2
    num_video_frames: int = 16
    needs_video_metadata: bool = False
    vllm_runner_kwargs: dict = field(default_factory=dict)
    marks: list = field(default_factory=list)


def get_compilation_config():
    """返回启用ViT CUDA graph的编译配置"""
    return {
        "cudagraph_mm_encoder": True, # 启用多模态编码器 CUDA graph
        "encoder_cudagraph_max_vision_items_per_batch": 1, # 每批最大视觉项数
        "encoder_cudagraph_max_frames_per_batch": 16, # 每批最大视频帧数
    }


@create_new_process_for_each_test()
def test_vit_cudagraph_image(model_id, vllm_runner, image_assets):
    config = MODEL_CONFIGS[model_id]
    if "image" not in config.modalities:
        pytest.skip(f"{model_id} does not support the image modality.")
    image_prompts = IMAGE_ASSETS.prompts({
        "stop_sign": config.image_prompt,
        "cherry_blossom": config.image_prompt,
    })
    images = [[asset.pil_image] for asset in image_assets]
    with vllm_runner(
        config.model,
        dtype=config.dtype,
        max_model_len=config.max_model_len,
        max_num_seqs=config.max_num_seqs,
        limit_mm_per_prompt={"image": 1},
        compilation_config=get_compilation_config(),
        **config.vllm_runner_kwargs,
    ) as vllm_model:
        outputs = vllm_model.generate_greedy(
            image_prompts, config.max_tokens, images=images
        )
        assert len(outputs) == 2
        output_ids, output_text = outputs[0]
        assert len(output_ids) > 0
        assert len(output_text) > 0
        assert isinstance(output_text, str)
```

评论区精华

在 review 中，gemini-code-assist 指出测试资产获取方式脆弱（通过索引依赖 fixture 顺序），建议使用 IMAGE_ASSETS.get_asset 显式获取。该建议未被作者采纳，但 PR 最终获批合并。这反映了测试代码可维护性的权衡。

风险与影响

低风险。未修改生产代码，仅新增测试和 CI 配置。风险包括测试依赖网络下载模型和 GPU 资源。对用户无影响，对团队增加了回归保障。

关联脉络

本 PR 与多模态 CUDA graph 优化相关，可视为该功能的测试配套。未来可扩展至更多模型，如 Gemma4 等。