

PR #45254 完整报告

vllm-project/vllm

[MM][CG] Support ViT full CUDA graph for Ernie-4.5-VL image inference

合并时间: 2026-08-06 13:04

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45254>

执行摘要

- 一句话: Ernie-4.5-VL 图像推理接入 ViT CUDA Graph, 编码器提速约 3.8 倍
- 推荐动作: 值得精读。该 PR 是 SupportsEncoderCudaGraph 协议的又一个落地案例, 展示了如何处理无法捕获的 host-side 逻辑 (resampler eager) 以及如何通过 prepare_encoder_metadata 解耦图捕获与 eager 路径。对于要接入 ViT CUDA Graph 的模型开发者有很强的参考价值, 尤其是固定缓冲区的填充策略与最坏情况 max_seqlen 的设计。

功能与动机

该 PR 是 ViT Full CUDA Graph tracker #38175 的一部分, 旨在将编码器 (ViT) 纳入 CUDA Graph 以消除图像编码阶段的内核启动与同步开销。作者引用 Qwen3-VL 的 SupportsEncoderCudaGraph 作为参考实现。PR body 中给出了明确收益: 在 H100 上 encoder_forward 均值从 eager 的 20.19ms 降至 CUDA Graph 的 5.33ms (约 3.79x 加速, P99 达 5.35x)。同时通过 AI2D 与 MMStar 基准验证了精度不受影响 (差异在 stderr 内), 并通过 50 张图的贪婪解码验证 94% 字节级一致输出。

实现拆解

1. 拆分 metadata 计算: 在 vllm/model_executor/models/ernie45_vl.py 的 Ernie4_5_VisionTransformer 中新增 prepare_encoder_metadata(), 将原 forward 内联的 rotary 嵌入、cu_seqlens、max_seqlen 计算迁出, 支持 max_batch_size 填充固定形状和 max_seqlen_override 最坏情况覆盖, 使 eager、捕获和回放三条路径共享同一套预计算逻辑。
2. 实现 SupportsEncoderCudaGraph 协议: Ernie4_5_VLMoeForConditionalGeneration 继承该协议, 实现 get_encoder_cudagraph_config (4 个 buffer key, out_hidden_size)、get_encoder_cudagraph_budget_range (64 到 min(max_num_batched_tokens, max_model_len))、get_encoder_cudagraph_item_specs、select_encoder_cudagraph_items (按累积 patch 偏移切分 pixel_values) 以及捕获输入与回放缓冲区的准备方法。捕获边界限定为 ViT, resampler 在 postprocess_encoder_output 与 encoder_eager_forward 中 eager 执行。
3. 测试配套: 在 tests/models/multimodal/generation/test_vit_cudagraph.py 中新增 ernie45_vl 配置, 使用 load_format="dummy" 与 dummy_hf_overrides 缩减为 1 层文本 + 1 层视觉的随机权重, 避免约 56 GiB 权重下载, 在任意 CI GPU 上验证捕获 / 回放正确

性。

4. 示例与文档：在 `examples/generate/multimodal/vision_language_offline.py` 的 `MODELS_SUPPORT_VIT_CUDA_GRAPH` 列表加入 `ernie45_vl`；在 `docs/design/cuda_graphs_multimodal.md` 表格中登记该模型（图像 CG 支持，视频暂不支持）。

关键文件：

- `vllm/model_executor/models/ernie45_vl.py`（模块 模型实现；类别 `source`；类型 `data-contract`；符号 `forward`, `prepare_encoder_metadata`, `get_encoder_cudagraph_config`, `get_encoder_cudagraph_budget_range`）：核心实现文件：新增 `prepare_encoder_metadata` 拆分 `metadata` 计算，并完整实现 `SupportsEncoderCudaGraph` 协议（配置、预算、item 规格、输入选择、捕获 / 回放缓冲）。捕获边限定 ViT，resampler 保持 `eager`。
- `tests/models/multimodal/generation/test_vit_cudagraph.py`（模块 测试；类别 `test`；类型 `test-coverage`；符号 `ernie45_vl_chat_template`）：新增 `ernie45_vl` 测试配置，使用 `dummy` 权重与 `chat template` 覆盖图像场景的捕获 / 回放，验证功能正确性。
- `examples/generate/multimodal/vision_language_offline.py`（模块 示例；类别 `source`；类型 `core-logic`）：将 `ernie45_vl` 加入 `MODELS_SUPPORT_VIT_CUDA_GRAPH` 列表，使用户能在离线示例中直接体验 ViT CUDA Graph 效果。
- `docs/design/cuda_graphs_multimodal.md`（模块 文档；类别 `docs`；类型 `documentation`）：在支持的模型能力表格中登记 Ernie-4.5-VL（图像 CG 支持，视频暂不支持），明确能力边界。

关键符号： `prepare_encoder_metadata`, `forward`, `get_encoder_cudagraph_config`, `get_encoder_cudagraph_budget_range`, `get_encoder_cudagraph_item_specs`, `select_encoder_cudagraph_items`, `prepare_encoder_cudagraph_capture_inputs`, `prepare_encoder_cudagraph_replay_buffers`

关键源码片段

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

新增 `ernie45_vl` 测试配置，使用 `dummy` 权重与 `chat template` 覆盖图像场景的捕获 / 回放，验证功能正确性。

```
def ernie45_vl_chat_template(content: str) -> str:
    # Ernie-4.5-VL 的对话模板，将图像占位符放在用户提问之后
    return (
        f"<|begin_of_sentence|>User: {content}"
        "Picture 1:<|IMAGE_START|><|image@placeholder|><|IMAGE_END|>\n"
        "Assistant: <think></think>"
    )

# 在 MODEL_CONFIGS 中新增的配置条目
def _ernie45_vl_config():
    return VitCudagraphTestConfig(
        model="baidu/ERNIE-4.5-VL-28B-A3B-PT",
```

```

# 仅图像：视频的 resampler 应用 temporal conv 改变输出 token 数，走 eager 路径
modalities=["image"],
image_prompt=ernie45_vl_chat_template("What is in this image?"),
# 缩减为随机权重的 1 层文本 + 1 层视觉，任意 CI GPU 可运行且跳过约 56 GiB 下载
vllm_runner_kwargs={
    "load_format": "dummy",
    "trust_remote_code": True,
    "revision": "refs/pr/17",
    "hf_overrides": partial(
        dummy_hf_overrides,
        model_arch="Ernie4_5_VLMoeForConditionalGeneration",
    ),
},
)

```

评论区精华

- shen-shanshan 在文档表格处提问：“As far as I know, this model supports video inference, does this PR support?” 作者回复确认视频推理本身支持，但视频路径在 resampler 中使用 use_temporal_conv 会改变输出 token 数，需要单独处理，故本 PR 仅图像，后续 PR 跟进。
- shen-shanshan 指出如果仅支持图像，某些协议方法（如 get_input_modality）不需要，并引用 PR #44484；作者回复已移除。
- shen-shanshan 询问 use_temporal_conv=False 时 resampler 是否可捕获（仅线性与激活层）。作者确认可行，但官方 ERNIE-4.5-VL-28B-A3B 配置 use_temporal_conv=True，因此 resampler 留在 eager 路径，包含在后续 PR 中。
- 视频推理支持范围 (design): 确认仅图像支持，视频留待后续 PR；文档如实标注。
- resampler 是否可纳入 CUDA Graph (design): 当前保留 eager 路径，后续 PR 可探索纳入图上。
- 图像专用协议方法冗余 (design): 已移除冗余方法。

风险与影响

- 风险：
 - 数值一致性：CUDA Graph 回放使用填充后的固定缓冲区，导致 3/50 图片出现单字翻转（接近平局），字节级不完全一致，但作者论证不影响模型输出质量。
 - 仅图像路径覆盖：视频输入继续走 eager，若调用方在视频场景误启用 cudagraph_mm_encoder，可能回退或产生未预期行为（需检查运行时是否按 modality 分流）。
 - prepare_encoder_metadata 被 eager 与图路径共用，若 max_seq_len_override 或 max_batch_size 使用不当（如覆盖值小于实际），会造成注意力掩码错误或越界；当前实现依赖调用方正确传入。
 - 依赖 SupportsEncoderCudaGraph 协议和 encoder_cudagraph_defs 的演进，后续框架改动需同步维护。

- 测试使用 dummy 权重与缩减层数，未验证真实权重下捕获 / 回放的数值稳定性（但作者提供了外部的精度基准）。
- 影响：
 - 对用户：使用 ERNIE-4.5-VL 图像推理的用户启用 cudagraph_mm_encoder 后可获得约 3.8x 编码器加速，GPU 空闲时间减少。
 - 对系统：新增一个模型接入 ViT CUDA Graph 协议，扩展了可捕获模型的覆盖范围；resampler 保持 eager，因此显存占用与 eager 路径接近，未增加图捕获池负担。
 - 对团队：确立了一种“拆分 metadata + 部分捕获”的模式，为后续视频路径与其他模型提供参考；文档登记了模型能力边界。
 - 风险标记：仅图像路径支持，resampler 保留 eager 路径，回放数值非逐位一致，依赖协议演进，测试使用 dummy 权重

关联脉络

- 暂无明显关联 PR