

VideoAuto-R1 模型推理与部署实战

1. 项目概述

1.1 核心创新

VideoAuto-R1 是一个采用 "reason-when-necessary" (按需推理) 策略的视频理解框架:

- 训练阶段:** 模型采用 "Thinking Once, Answering Twice" (思考一次, 回答两次) 范式
 - 首先生成初始答案
 - 然后进行推理思考
 - 最后输出审核后的答案
- 推理阶段:** 模型通过初始答案的置信度分数决定是否继续推理 (Early Exit 机制)

1.2 模型系列

模型	基座模型	参数量	上下文长度	HuggingFace
VideoAuto-R1-Qwen2.5-VL-7B	Qwen2.5-VL-7B-Instruct	7B	16K	IVUL-KAUST/VideoAuto-R1-Qwen2.5-VL-7B
VideoAuto-R1-Qwen3-VL-8B	Qwen3-VL-8B-Instruct	8B	128K	IVUL-KAUST/VideoAuto-R1-Qwen3-VL-8B

2. 环境配置

2.1 系统要求

- 操作系统:** Linux (Ubuntu 22.04)
- Python:** 3.12
- CUDA:** 12.8
- 显存要求:** 推理: 32GB (单卡)

2.2 创建环境

```
# 获取项目 (使用我提供的升级版代码)
```

```
cd VideoAuto-R1

# 创建 conda 环境
conda create -n videoauto-r1 python=3.12
conda init bash && source /root/.bashrc
conda activate videoauto-r1

# 安装依赖
pip install -r requirements.txt

# 安装 Flash Attention (加速推理)
# 先下载文件，再安装
wget https://github.com/Dao-AI-Lab/flash-attention/releases/download/v2.8.3/flash_attn-2.8.3+cu12torch2.8cxx11abiTRUE-cp312-cp312-linux_x86_64.whl

pip install flash_attn-2.8.3+cu12torch2.8cxx11abiTRUE-cp312-cp312-linux_x86_64.whl

# 安装 huggingface_hub
pip3 install --upgrade huggingface_hub
```

3. 模型下载

```
# 禁用 hf-xet
export HF_HUB_DISABLE_XET=1

# 下载模型
hf download IVUL-KAUST/VideoAuto-R1-Qwen3-VL-8B --local-dir
./models/VideoAuto-R1-Qwen3-VL-8B
```

4. 服务部署

模式	说明	适用场景
first	始终使用第一个答案	快速响应，低延迟场景
second	始终使用第二个答案（经过推理）	追求高准确率
auto	根据置信度自动决定	平衡速度与准确率（推荐）

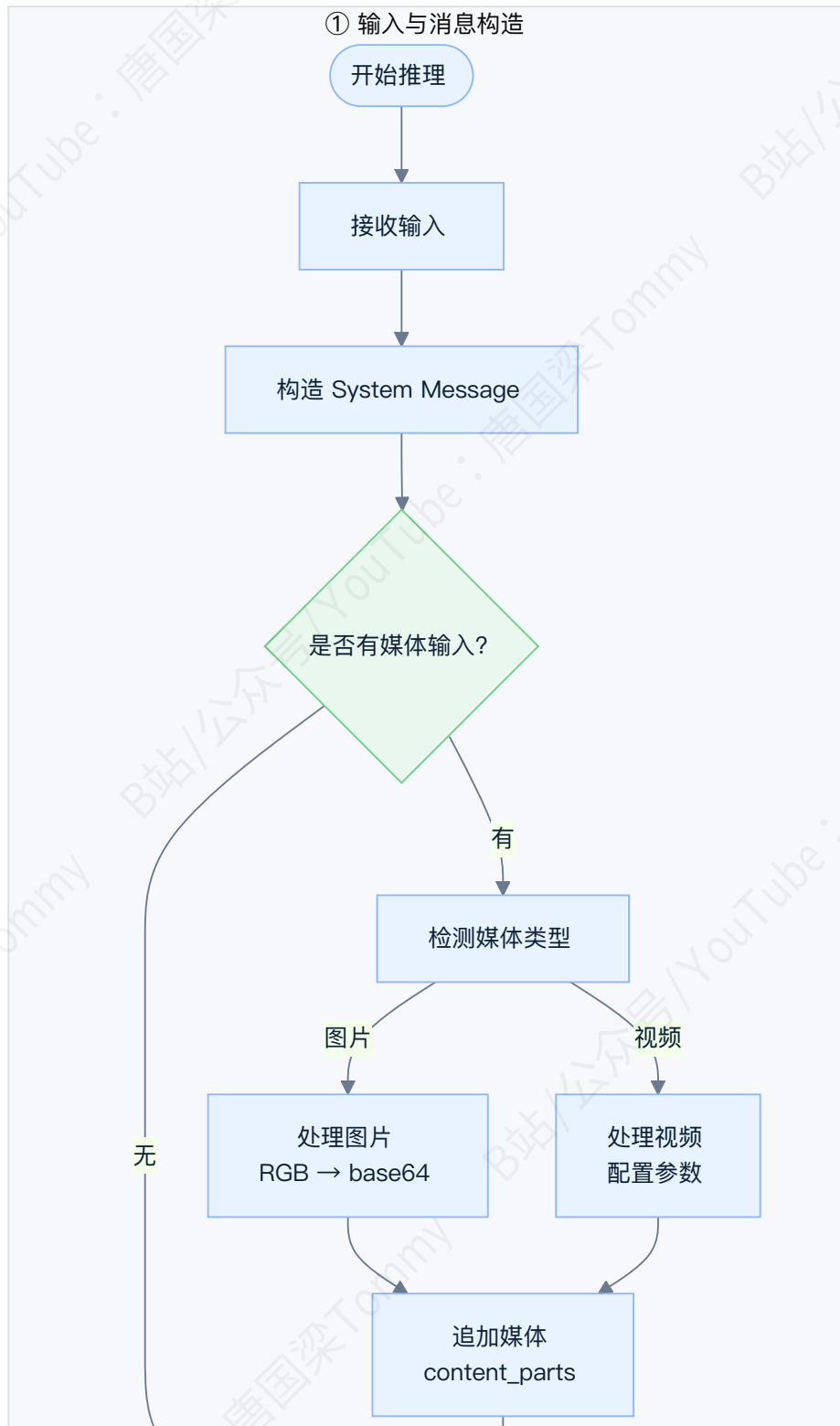
方案一：Web Demo 部署

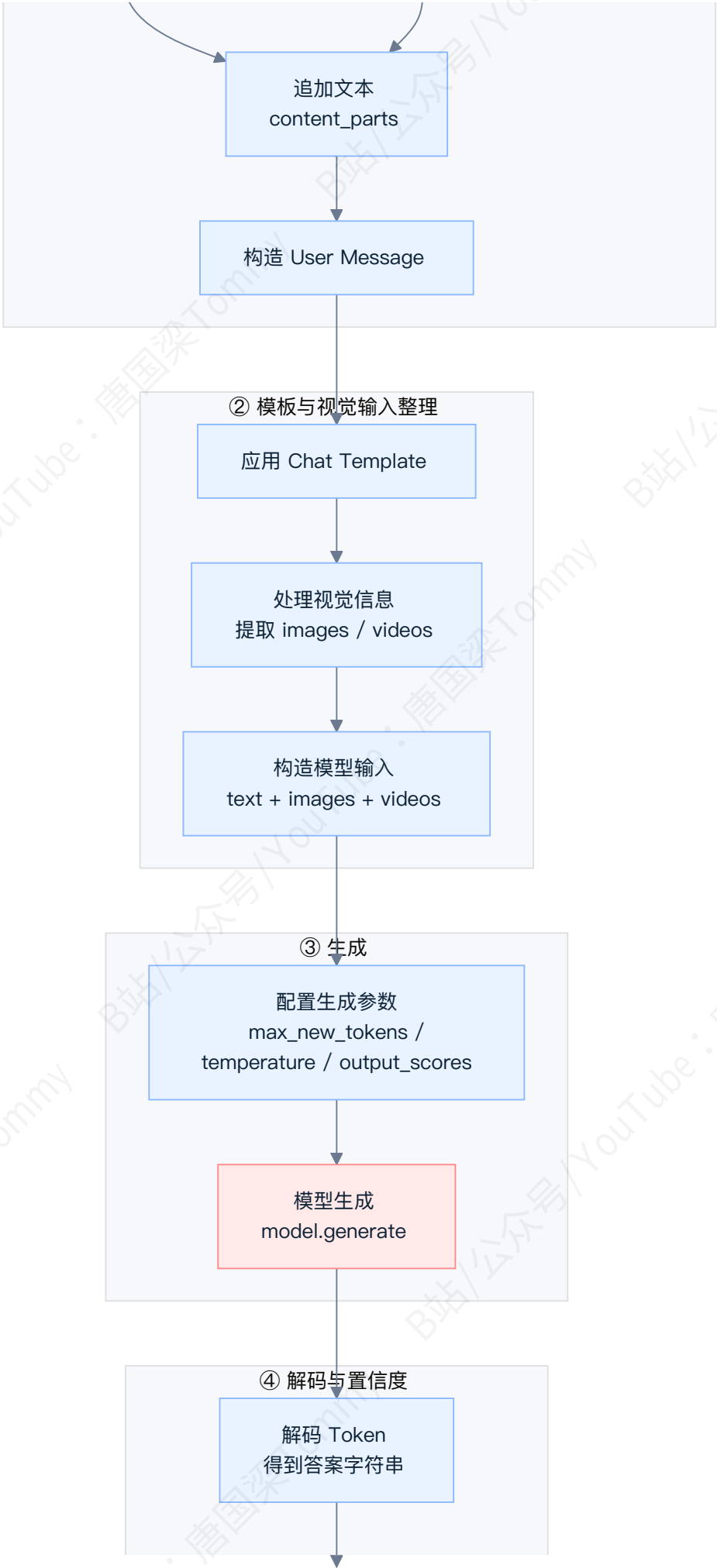
项目提供了完整的 Gradio 演示应用：

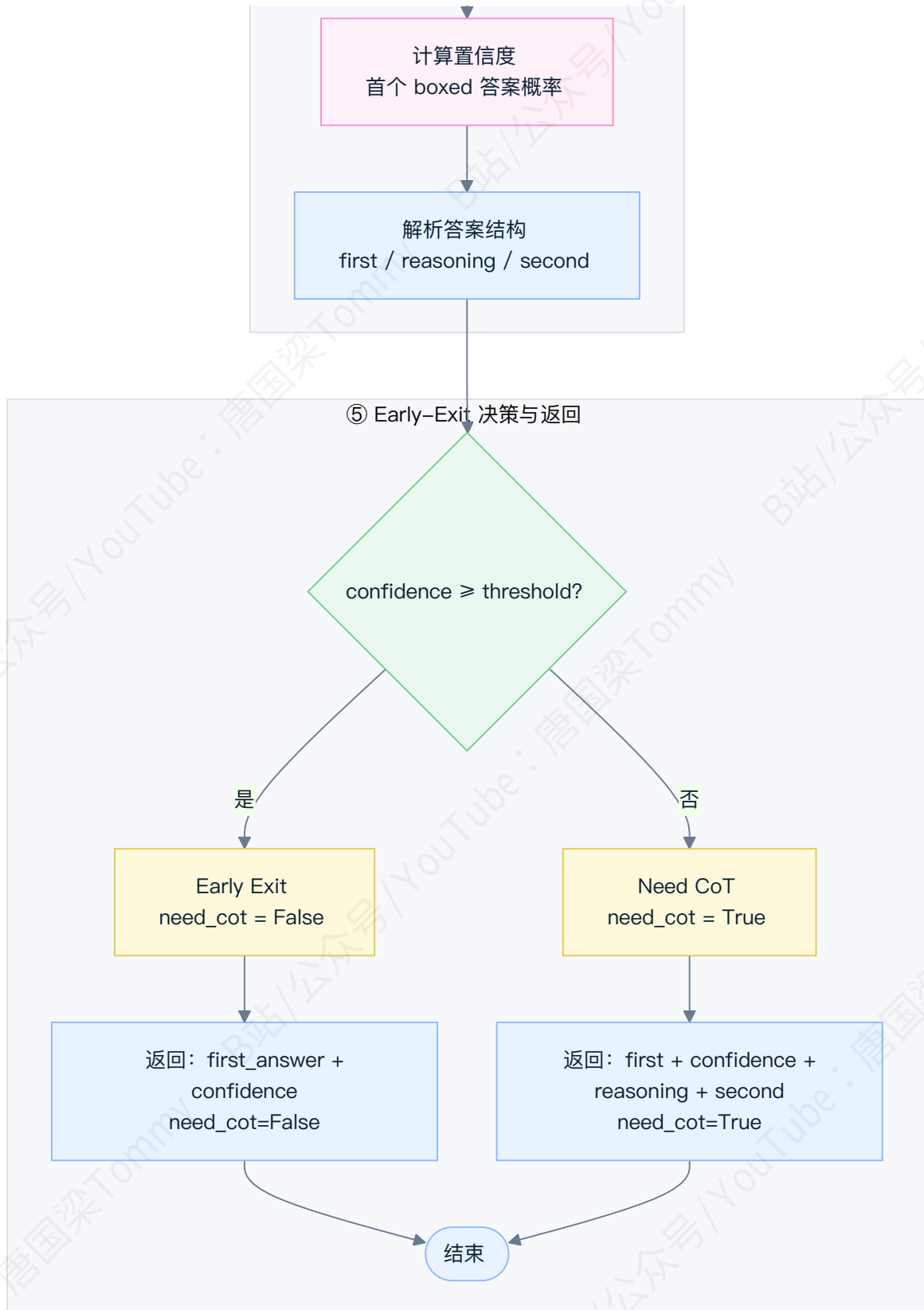
```
cd demo
python app.py
```

Demo 核心功能:

- 支持图像和视频上传
- 实时流式输出
- 可调节 Early Exit 阈值
- 显示置信度和推理过程







方案二：模型服务 API 部署

服务部署

```
uvicorn api_server:app --host 0.0.0.0 --port 8000
```

服务测试

```
uvicorn api_server:app --host 0.0.0.0 --port 8000
```

1. 纯文本

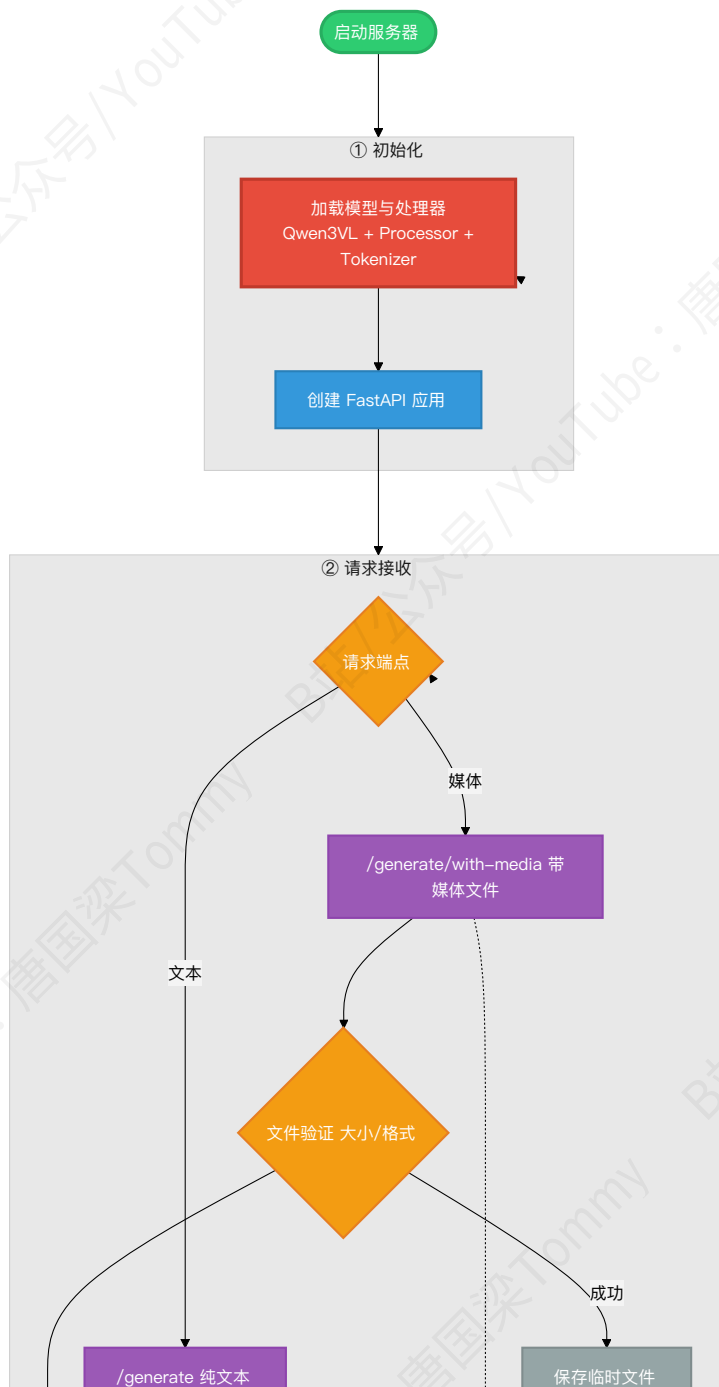
```
python test_api.py --text "请介绍一下火星"
```

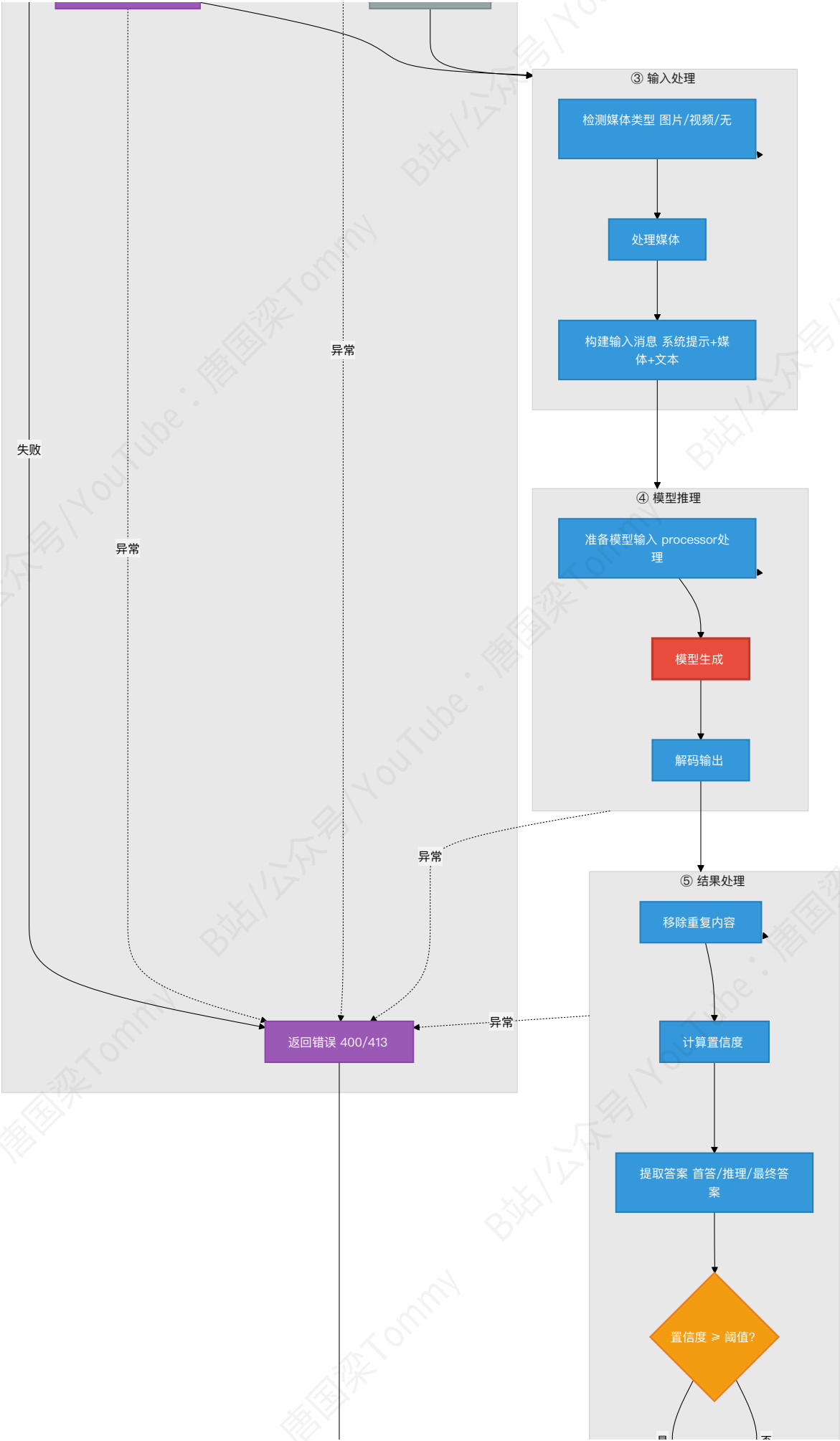
2. 纯图片

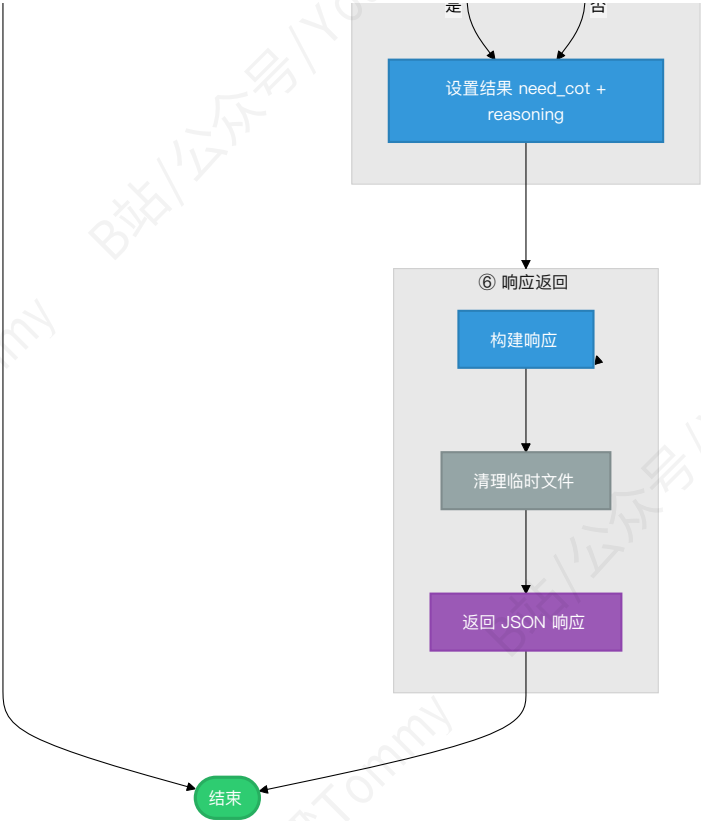
```
python test_api.py --image "./assets/test-1.jpg" "用中文详细描述一下图片中的内容"
```

3. 纯视频

```
python test_api.py --video "./assets/test-2.mp4" "用中文详细描述一下视频中的内容"
```







AI进阶学习

B站课堂：

<https://space.bilibili.com/474347248/pugv>

<https://space.bilibili.com/3546748815411393/pugv>

官方网站： <https://www.tgltommy.com/> (国内访问需科学上网)

TGL

AI

Tommy

Browse products

Home

Contact

About Me

Sign up

Log in

Categories

Premium Courses

Public Courses

DeepSeek

系统实战

架构·算法·工程·应用

唐国梁Tommy | TGLTommy

DeepSeek系统实战：架构·算法·工程·应用

这是一门面向真实工程落地的大模型系统课程，围绕 DeepSeek 的技术路线，构建从“可控部署与应用搭建”到“核心机制与对齐算法”，再到“系统...

Course · By 唐国梁Tommy

Learn more

多模态大模型

前沿算法与实战应用

第一季 图文与视频理解

唐国梁Tommy

多模态大模型 前沿算法与实战应用（第一季：图文与视频理解）

本课程是一门面向多模态大模型前沿技术的系列课程，深入探讨了当前多模态大模型的核心技术与最新进展。课程覆盖了从基础概念到高级算法实现...

Course · By 唐国梁Tommy

Learn more

大模型Llama架构

从理论到实战

核心技术与底层原理全解析

唐国梁Tommy

大模型Llama架构：从理论到实战

本课程对 Llama 架构的核心模块进行了深入解析，涵盖算法原理和技术实现。你将学习如何对 Llama模型进行自定义微调、以及本地化模型推...

Course · By 唐国梁Tommy

Learn more

大模型公开课

RAG/Agent/多模态

唐国梁Tommy

大模型公开课（持续更新）

欢迎你加入我的免费公开课~我将在这里将持续分享最新AI前沿技术内容，以简明、实用的方式带你快速了解当下热门的人工智能方法与应用。课程...

Course · By 唐国梁Tommy

Learn more

深入LLM与RAG

原理、实现与应用

大语言模型LLM与检索增强生成RAG理论与实战

唐国梁Tommy

深入LLM与RAG 原理、实现与应用

本课程全面聚焦于检索增强生成 (RAG) 的理论与实践，深入讲解如何结合主流的大语言模型 (LLMs) 来实现智能信息检索和生成。课程从项目环境配...

Course · By 唐国梁Tommy

Learn more

如果你想系统学习多模态大模型相关技术，推荐你学习我的精品课程《多模态大模型 前沿算法与实战应用 第一季》，课程覆盖通用/领域多模态架构、多模态数据范式、训练流程、评估与幻觉分析，并配套多项目实战：LLaVA、LLaVA-NeXT、Qwen2/3-VL、InternLM-XComposer（IXC）、TimeSearch-R等，包含算法讲解、服务部署、核心源码Debug解析。本课程近期正在更新中，欢迎加入学习。

《多模态大模型 前沿算法与实战应用 第一季》课程链接：

B站：<https://www.bilibili.com/cheese/play/ss33184>

官网（国内访问需科学上网）：<https://www.tgltommy.com/p/multimodal-season-1>

© 2026 TGLTommy 原创出品 · All Rights Reserved

B站/公众号/YouTube: 唐国梁Tommy

欢迎关注微信公众号



官方网站: TGLTommy.com(需科学上网)