Xinference 完整落地方案（RTX 4060 8G 专属版）

Xinference 完整落地方案（RTX 4060 8G 专属版）

文档说明：本方案为 Xinference 全场景通用部署手册，涵盖单机运行、远程多机访问、OpenClaw 对接、故障排查、系统优化全流程，适配 Ubuntu 系统、RTX 4060 8G 显卡，所有命令可直接复制执行，可直接保存为 Word/PDF 备用。

一、日常运行（优先级最高，每次使用直接参考）

1. 核心启动命令

- 本地单机启动

```
代码块
1  source ~/xin-env/bin/activate
2  xinference-local
```

- 开放网络启动（支持远程 / 多机访问，必用）

```
代码块
1  source ~/xin-env/bin/activate
2  xinference-local --host 0.0.0.0
```

2. 访问地址

- 本机访问：http://localhost:9997
- 同局域网其他设备访问：http:// 主机内网 IP:9997

3. 停止服务

在启动服务的终端，按下 `Ctrl + C` 即可终止服务，关闭终端会直接停止程序。

二、固定版本选择（无坑稳定版）

推荐安装版本：**xinference == 2.5.0**

- 完美适配 NVIDIA RTX 4060 显卡，CUDA 加速无兼容问题
- 支持 GGUF 格式模型，国内 ModelScope 源高速下载
- 运行稳定，无冗余依赖冲突，适配个人本地部署

三、安装前置准备（初次部署必做）

安装 Ubuntu 系统基础依赖，保障后续安装流程顺畅：

代码块

```
sudo apt update
sudo apt install -y python3-full python3-pip python3-venv
```

四、虚拟环境 + 软件安装（清华国内源）

1. 创建纯净虚拟环境

代码块

```
rm -rf ~/xin-env
python3 -m venv ~/xin-env
```

2. 激活虚拟环境

代码块

```
source ~/xin-env/bin/activate
```

终端前缀显示 `(xin-env)` 即为激活成功。

3. 安装 Xinference

代码块

```
pip install xinference==2.5.0 -i https://pypi.tuna.tsinghua.edu.cn/simple
```

五、显卡驱动与 CUDA 配置（留档备用）

1. 驱动与 CUDA 安装

```
sudo apt install -y nvidia-driver-550 nvidia-cuda-toolkit
```

2. 重启生效配置

代码块

```
sudo reboot
```

3. 验证显卡状态

代码块

```
nvidia-smi
```

正常输出需显示：RTX 4060 显卡、驱动版本≥550、CUDA 版本≥13.0

六、模型运行参数配置（8G 显存专属）

启动模型时，按照以下参数配置，兼顾速度与效果：

配置项	推荐值	配置说明
模型引擎	llama.cpp	轻量高效，完美适配 GGUF 模型，显卡加速最优
模型格式	ggufv2	主流通用格式，兼容性拉满
量化等级	Q4_K_M	8G 显存专属，平衡效果与显存占用
GPU 层数	99	全模型层加载至显卡，运行速度拉满
GPU 索引	0	单显卡默认配置
副本数	1	单实例运行，不浪费硬件资源
思考模式	开启	提升模型推理逻辑性
下载源	modelscope	国内阿里源，下载无卡顿
多模态投影器	留空	纯文本模型无需配置，避免依赖报错

七、显存与模型配置对照表

模型规格	量化等级	显存占用	适配性
7B/9B 通用模型	Q4_K_M	5-6GB	RTX 4060 8G 完美流畅运行
8B 系列模型	Q4_K_M	4-5GB	无压力运行，可后台常驻

八、远程访问与防火墙配置

1. 远程访问前提

启动命令必须添加 `--host 0.0.0.0`，开放网络端口

2. 获取主机内网 IP

代码块

```
hostname -I
```

3. 防火墙放行端口

代码块

```
sudo ufw allow 9997
sudo ufw reload
```

九、OpenClaw 对接配置（新增）

1. 功能说明

OpenClaw 作为本地模型统一调度客户端，可直接对接 Xinference API，实现一站式模型管理、对话交互、多模型切换。

2. 对接步骤

- 确保 Xinference 已用 `--host 0.0.0.0` 启动，网络正常
- 打开 OpenClaw，进入模型 API 配置页面
- 填写配置信息：
  - API 地址：`http://Xinference主机内网IP:9997/v1`

- API Key：任意填写（本地部署无需验证，如 xxx）
- 模型名称：Xinference 中运行的模型名称（如 qwen3.5）

- 保存配置，即可在 OpenClaw 中直接调用 Xinference 模型

3. 注意事项

OpenClaw 与 Xinference 需处于同一局域网，确保网络互通。

十、模型推理调用方式

1. Web 界面调用

直接访问 Xinference 管理页面，点击模型对应「对话」按钮，即可直接交互。

2. API 通用调用

适配各类支持 OpenAI 协议的客户端 / 程序，调用参数：

- API 地址：`http://主机内网IP:9997/v1`
- API Key：任意填写
  调用示例：

代码块

```
curl http://主机内网IP:9997/v1/chat/completions \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer xxx" \
  -d '{
    "model": "qwen3.5",
    "messages": [{"role": "user", "content": "你好"}],
    "temperature": 0.7
  }'
```

十一、常见问题排查

1. 模型下载卡顿 / 卡死：按 `Ctrl + C` 重启服务，重新启动模型；或手动下载 GGUF 文件，本地路径导入
2. 远程设备无法访问：检查启动命令、防火墙是否放行 9997 端口、内网 IP 是否正确
3. 显存占用过高：关闭多余模型实例，确认 GPU 层数设为 99
4. 依赖安装卡住：手动执行 `pip install xllamacpp -i https://pypi.tuna.tsinghua.edu.cn/simple`

十二、系统优化建议

1. 关闭电脑自动休眠、系统自动更新，避免服务中断
2. 优先使用 ModelScope 国内源下载模型，减少网络波动
3. 定期清理缓存目录：`~/.xinference/cache/`，释放磁盘空间
4. 运行模型时，关闭无关后台程序，释放显存与 CPU 资源

十三、核心命令速记

代码块

```
# 激活虚拟环境
source ~/xin-env/bin/activate
# 开放网络启动
xinference-local --host 0.0.0.0
# 查看内网IP
hostname -I
# 放行9997端口
sudo ufw allow 9997
# 验证显卡
nvidia-smi
# 关闭指定模型
xinference model terminate --model-id 模型ID
```

文档保存方法

1. 全选本文档内容，复制粘贴至 Word，调整格式后保存为.docx 文件
2. 如需 PDF 格式，通过 Word「另存为 PDF」或浏览器打印功能导出

（注：文档部分内容可能由 AI 生成）