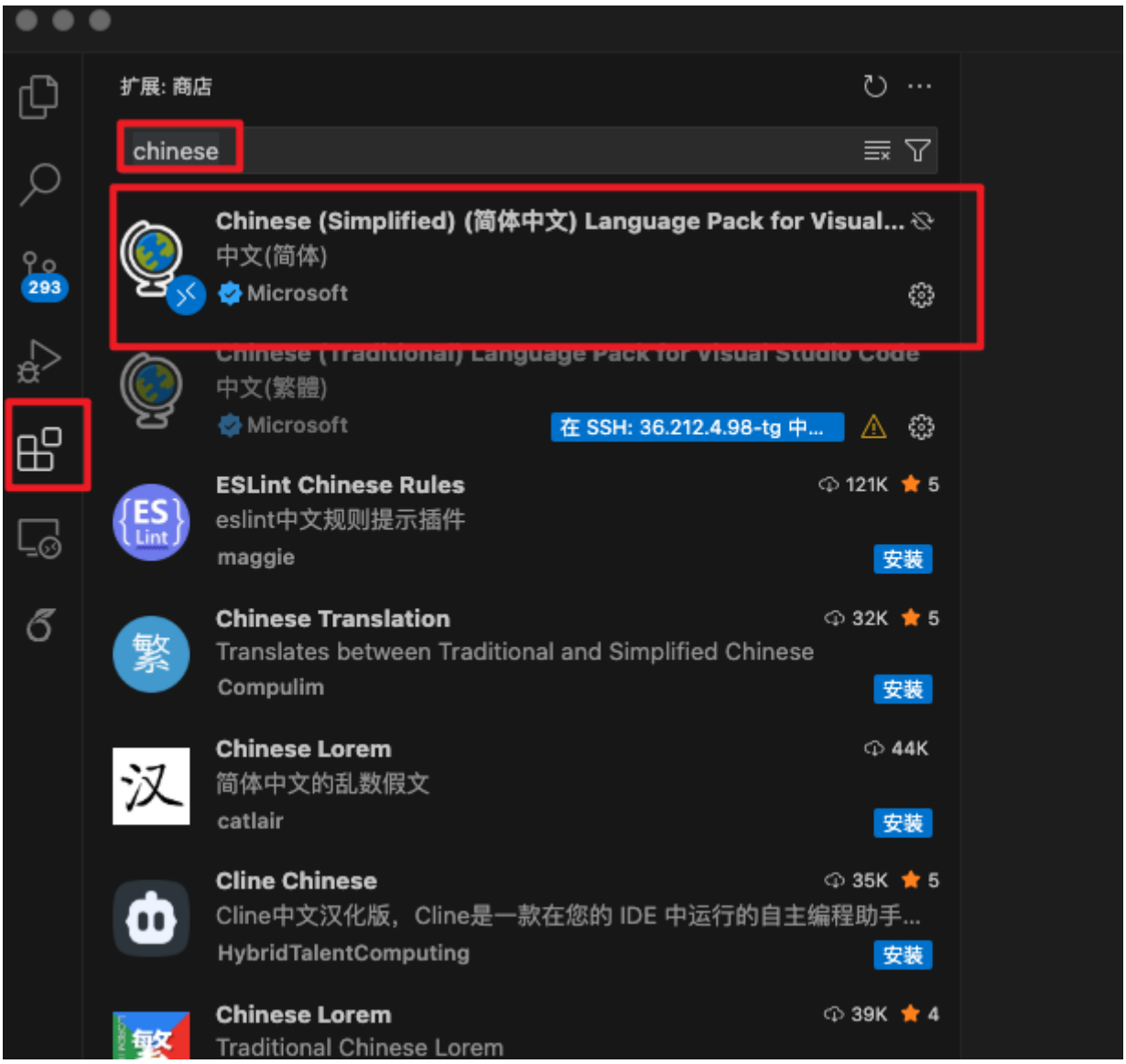
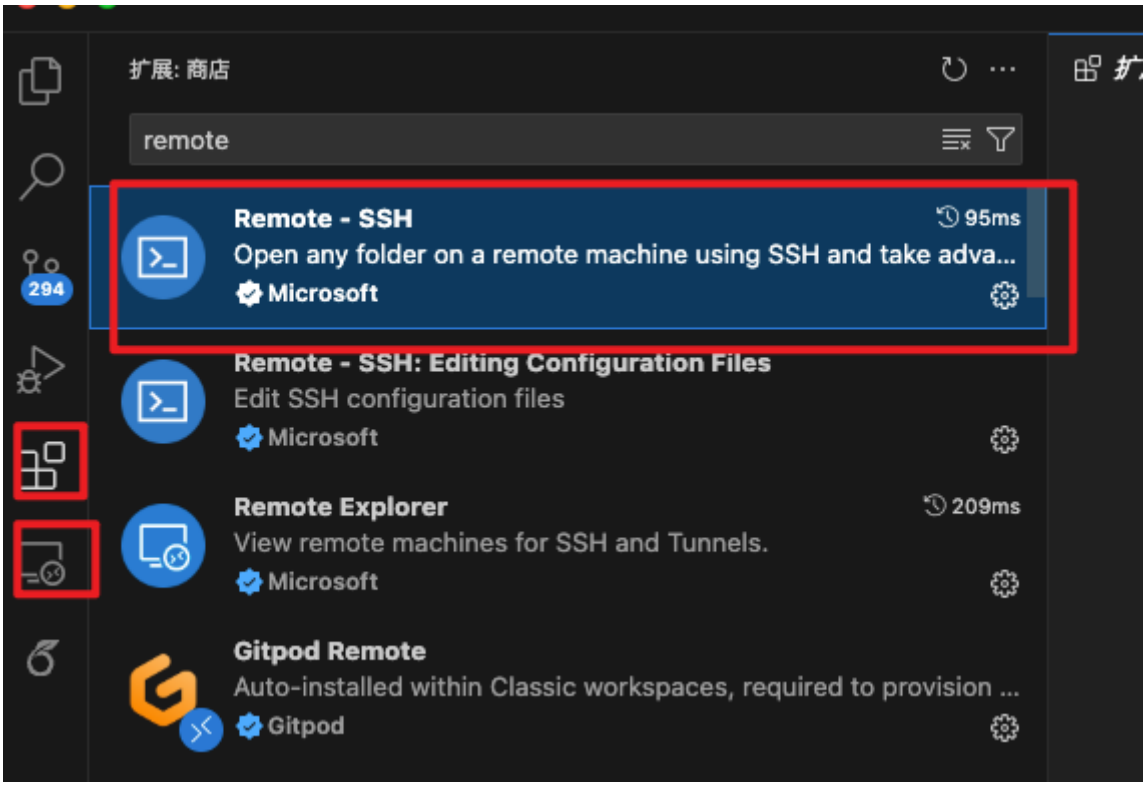
教程详细版

视频版地址： 链接: https://pan.baidu.com/s/1GTNBCrZ5hBw4w2CYj2jMeA?pwd=73c8 提取码: 73c8

vscode

vscode插件安装：chinese、remote、python、pylance、python debugger、Python Environment Manager

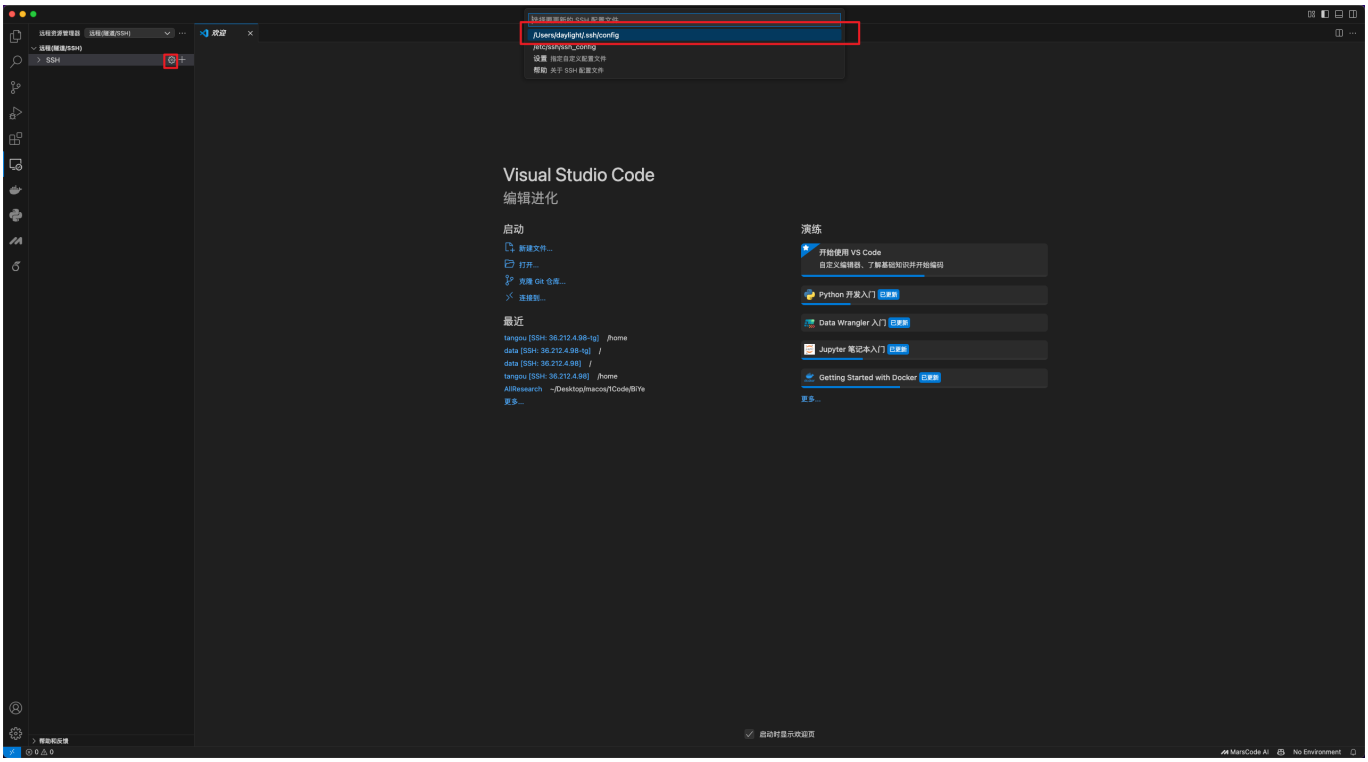




服务器连接

大家查看群文件自己的user和密码

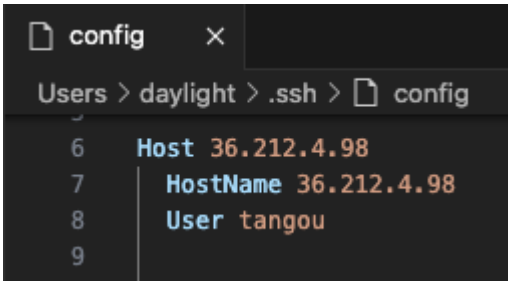
vscode连接:



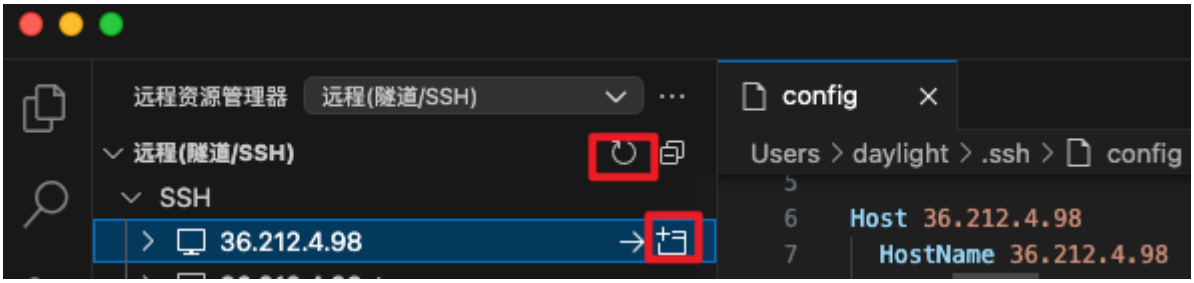
1. 复制到文件里面去

```
Host 36.212.4.98
  HostName 36.212.4.98
```

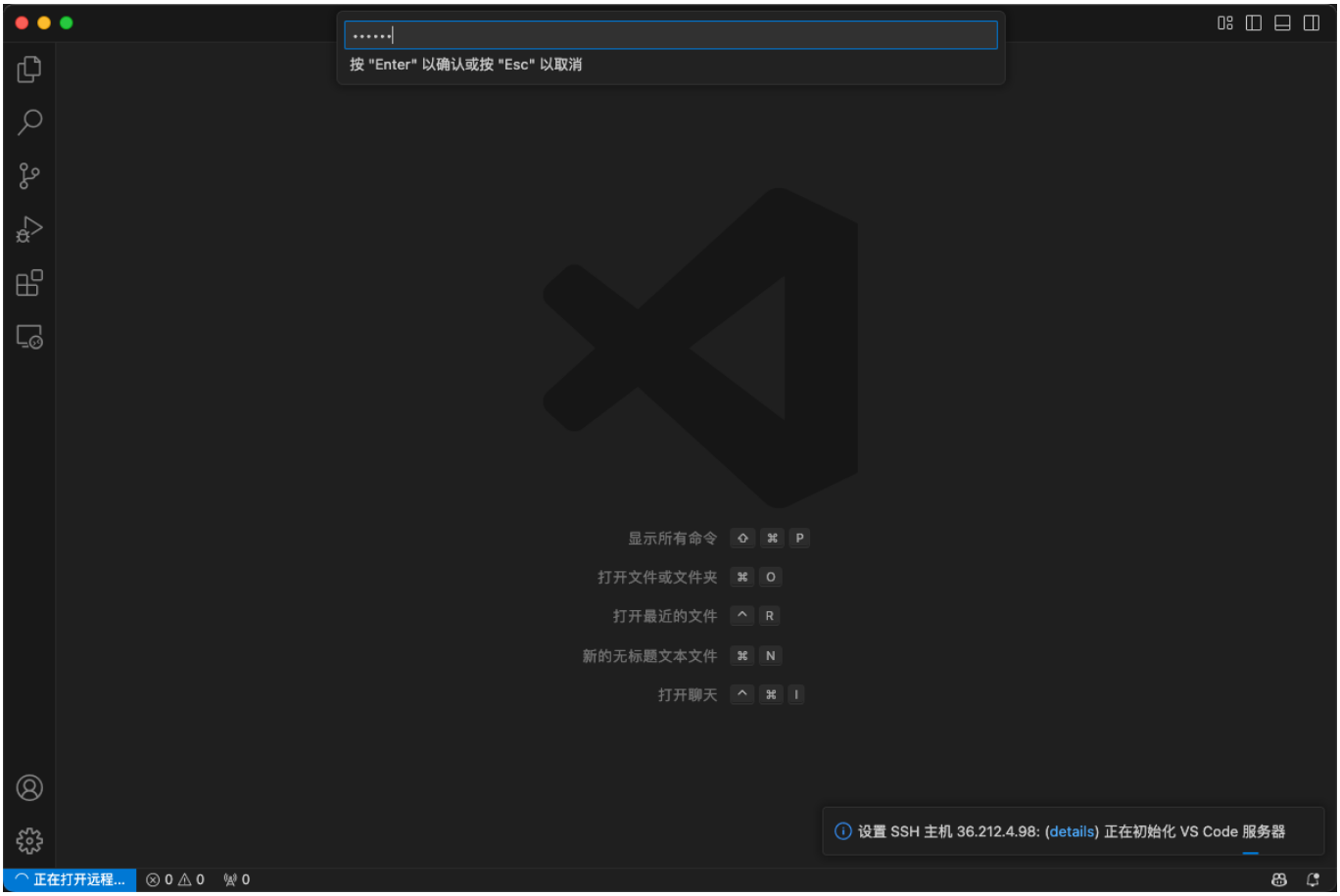
```
    User tangou
```



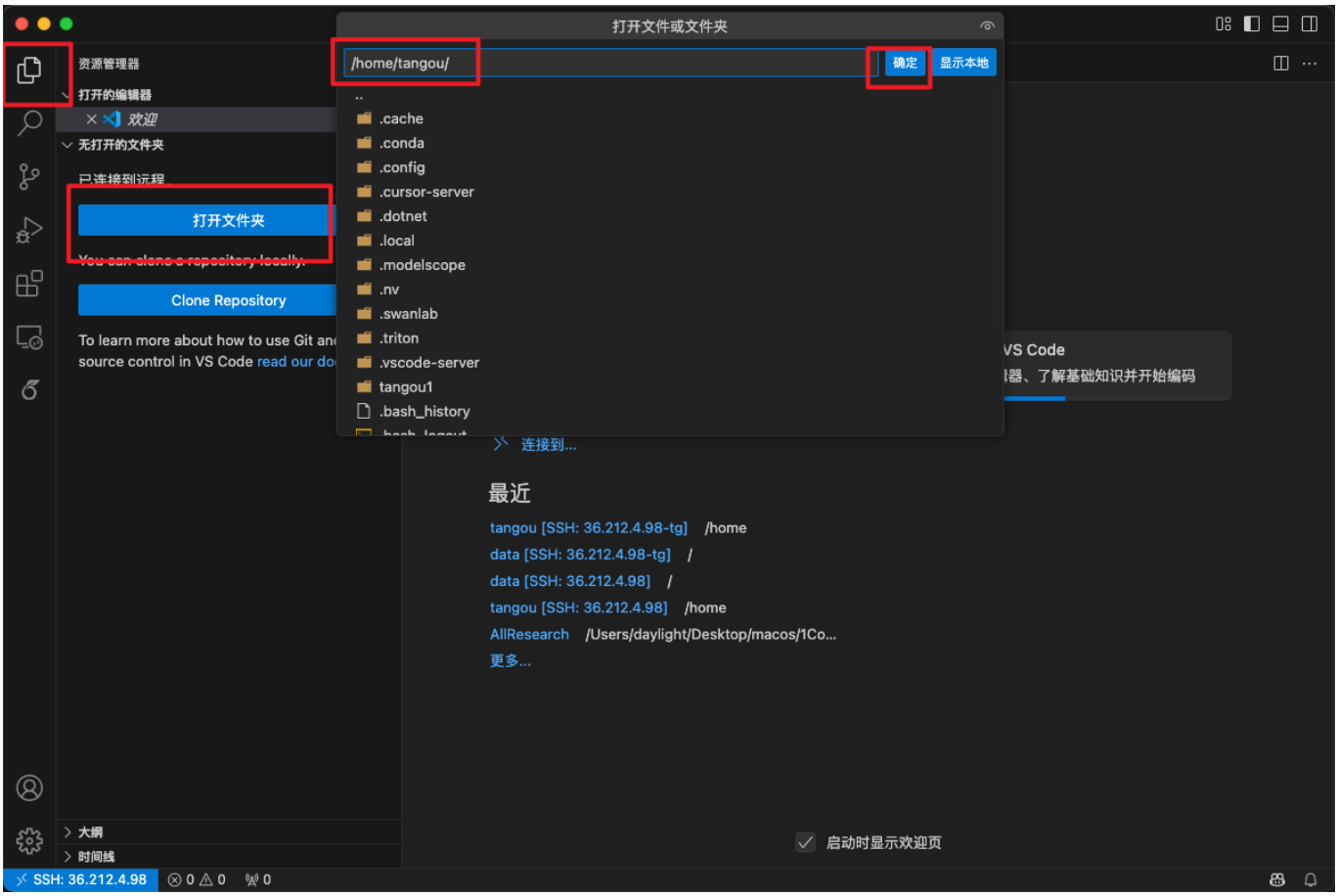
2. 点刷新，再打开文件



3. 输入密码123456，回车

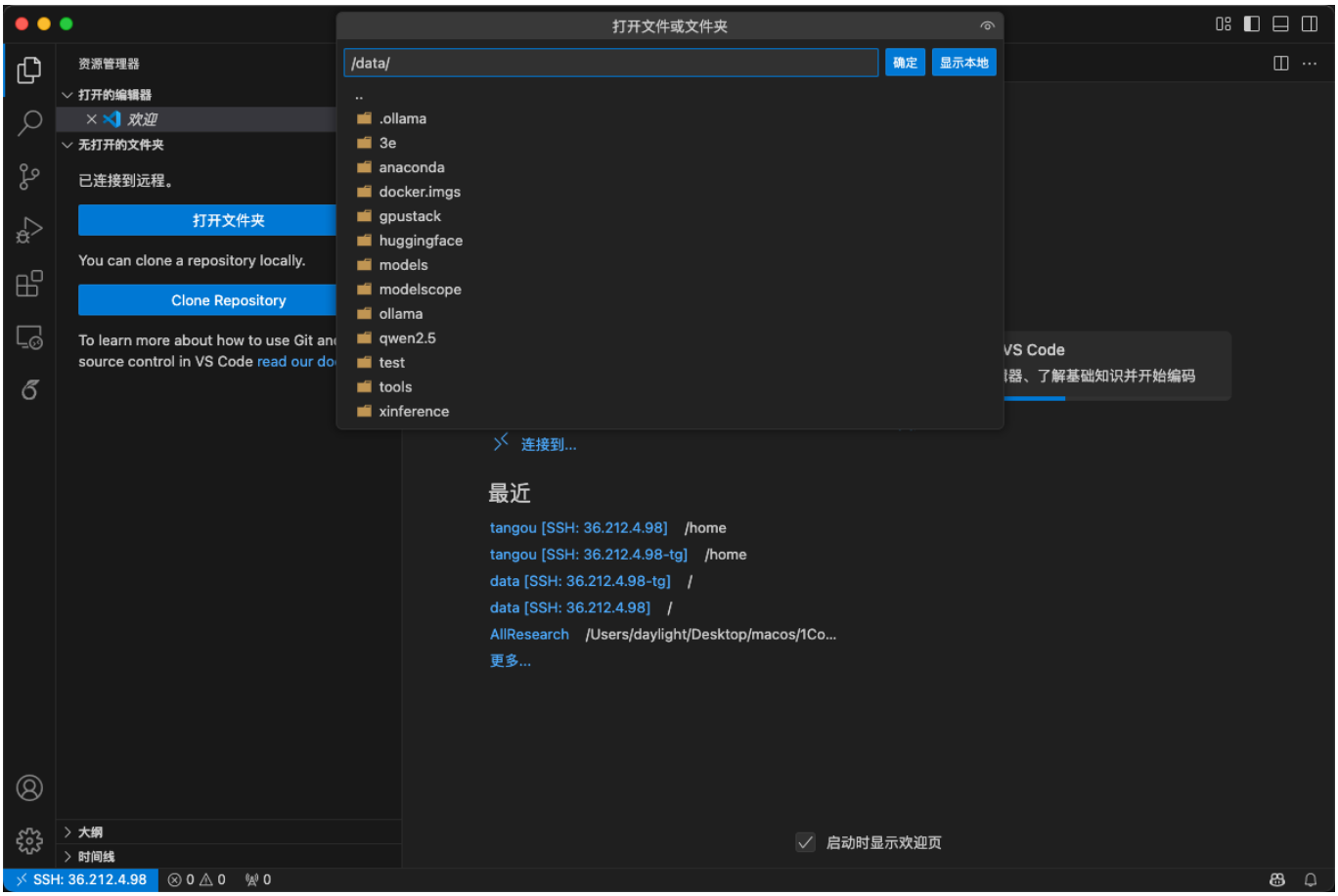


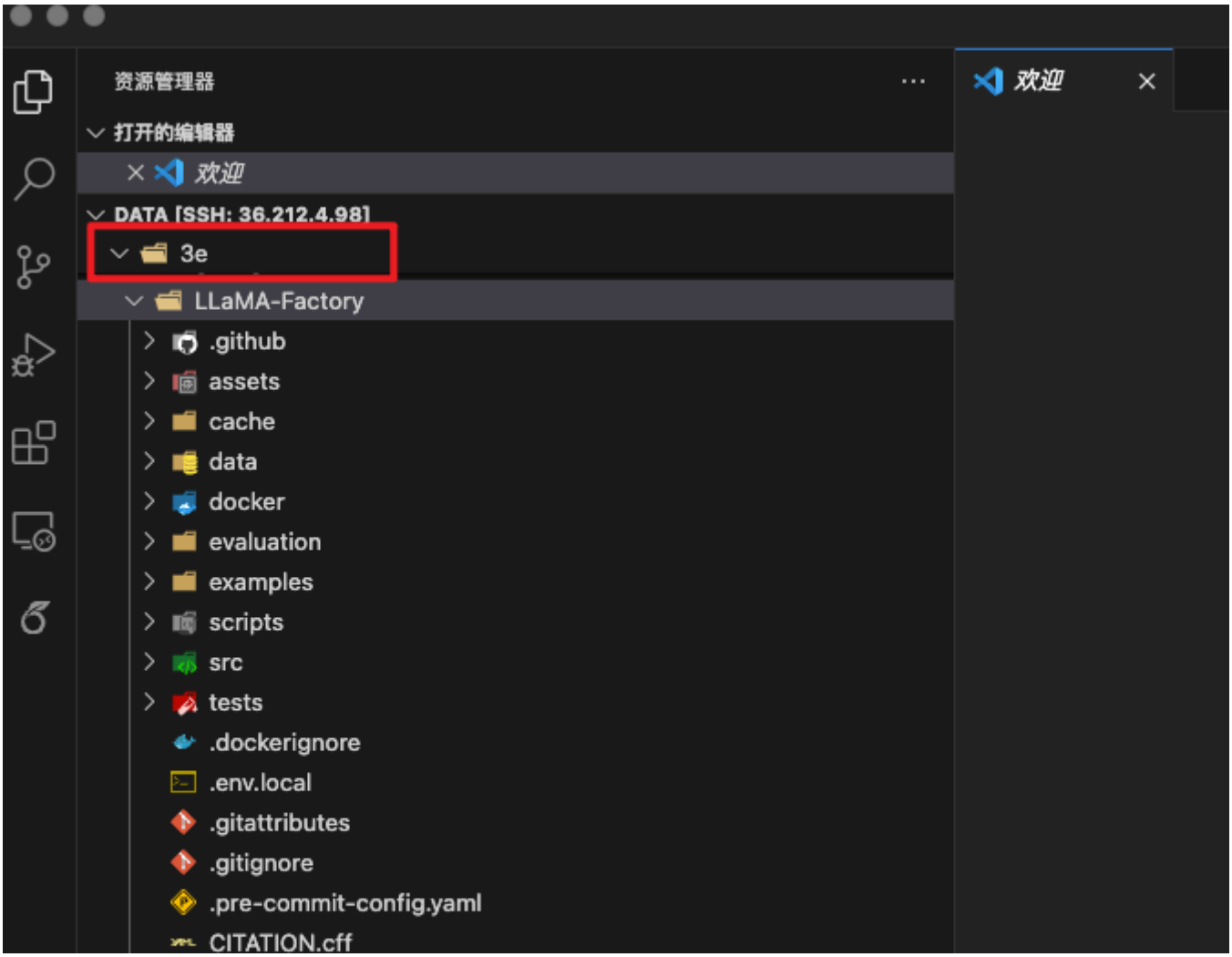
4. 进入

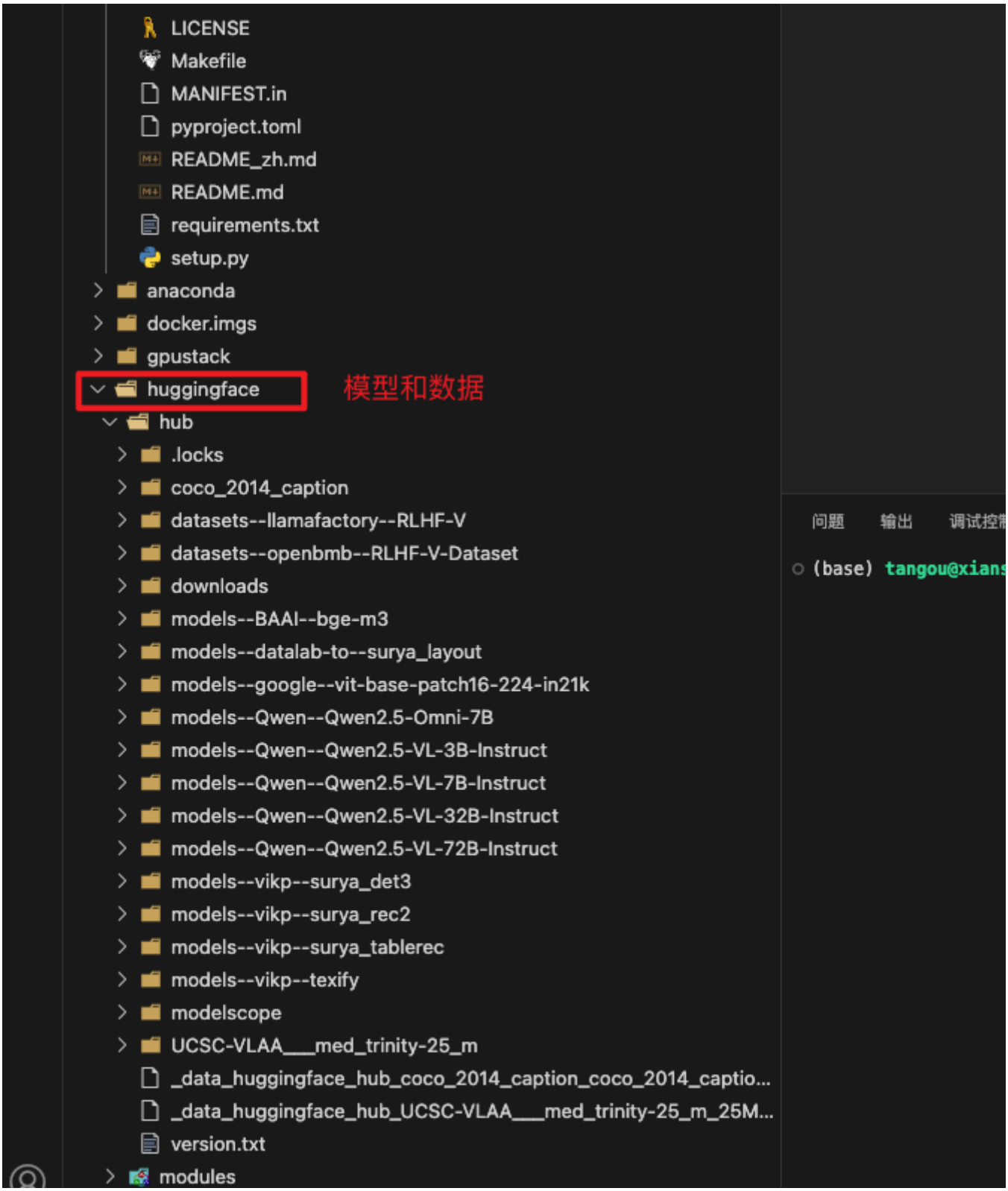

打开文件或文件夹
/home/tangou/
确定
显示本地
.cache
.conda
.config
.cursor-server
.dotnet
.local
.modelscope
.nv
.swanlab
.triton
.vscode-server
tangou1
.bash_history
资源管理器
打开的编辑器
欢迎
无打开的文件夹
打开文件夹
Clone Repository
连接到...
最近
tangou [SSH: 36.212.4.98-tg] /home
data [SSH: 36.212.4.98-tg] /
data [SSH: 36.212.4.98] /
tangou [SSH: 36.212.4.98] /home
AllResearch /Users/daylight/Desktop/macos/1Co...
更多...
启动时显示欢迎页
大纲
时间线
SSH: 36.212.4.98


打开的编辑器
TANGOU [SSH: 36.212.4.98-TG]
建文件
建文件夹
.cache
.conda
.config
.cursor-server
.dotnet
.local
.modelscope
.nv
.swanlab
.triton
.vscode-server
tangou1
tangou2
建个自己的文件夹
.bash_history
.bash_logout
.bashrc
.gitconfig
.lesshst
.profile
.wget-hsts

5. 回到第2步，再去开一个目录。这是共享模型数据的目录



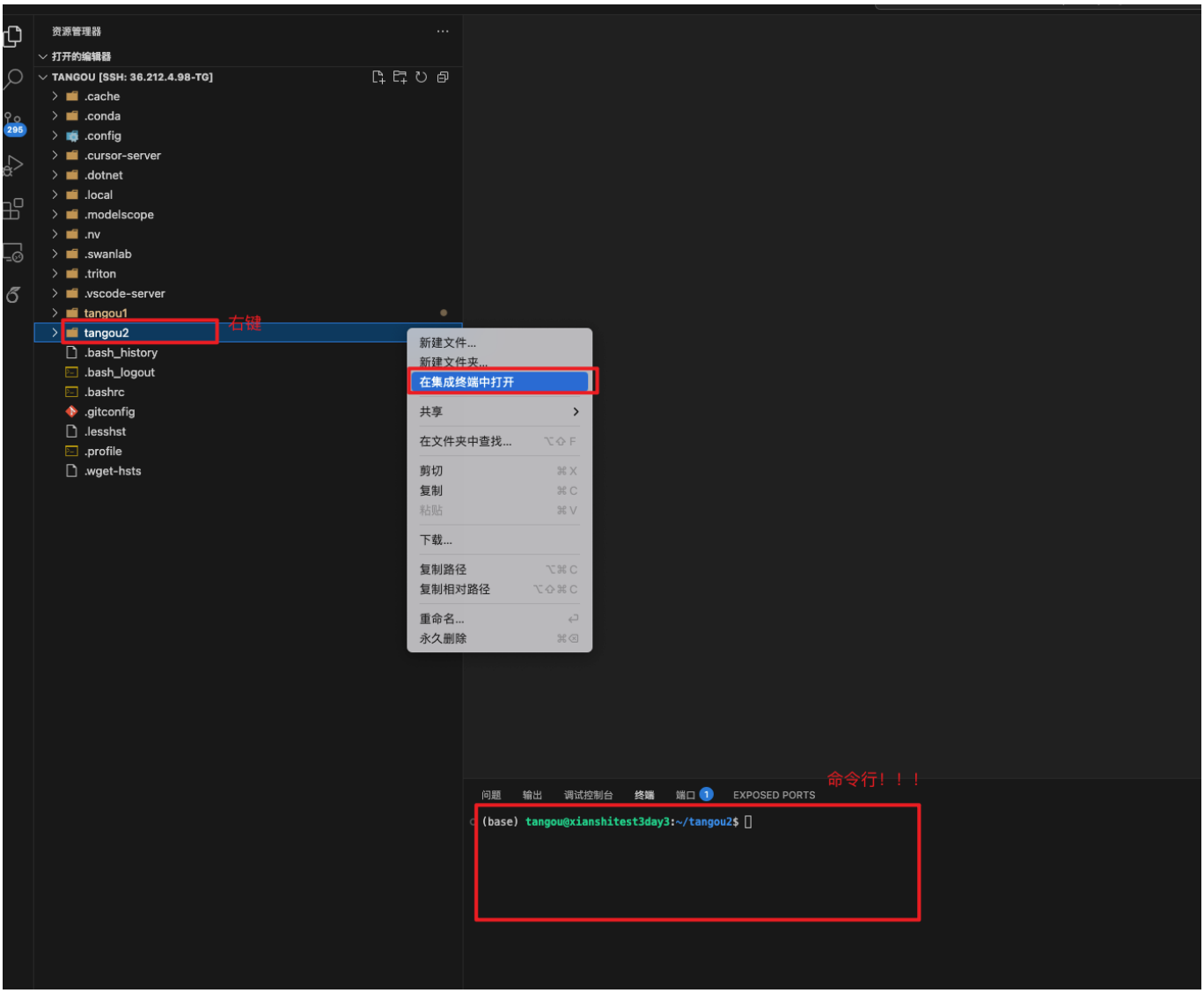




微调预备

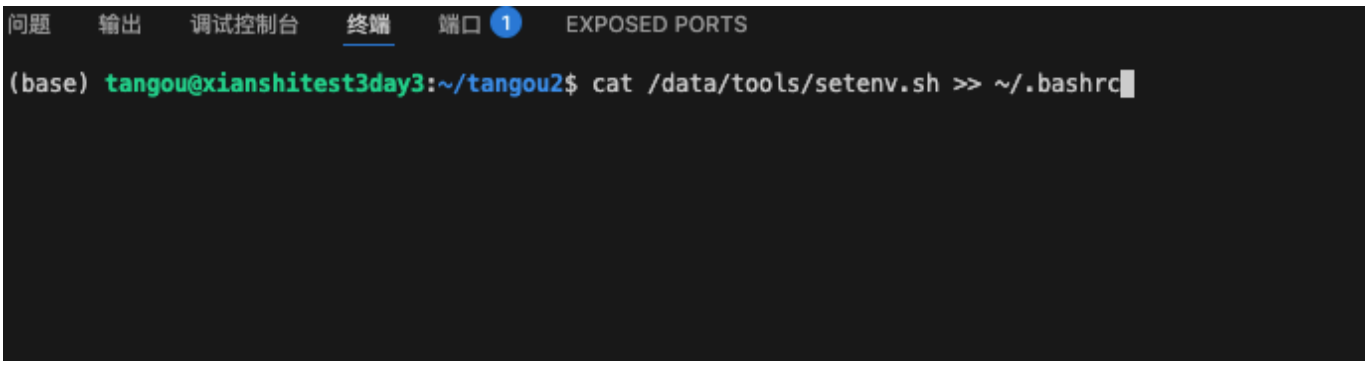
- 环境变量：conda、ollama

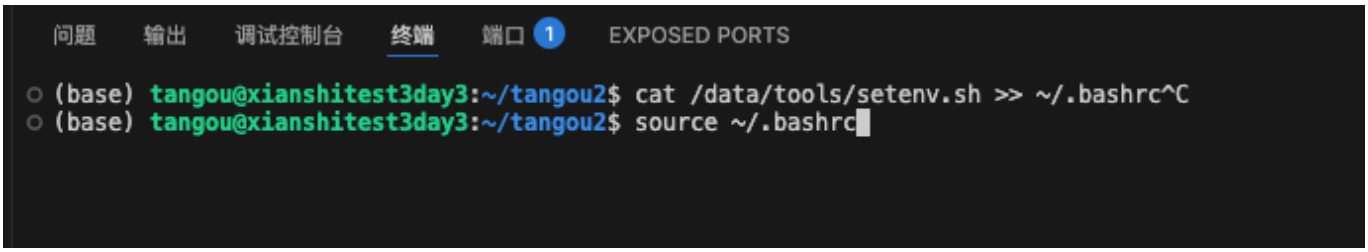
1. 打开终端



2. 命令行运行，复制命令过去回车，运行。下面截图我之前运行过了，没运行。

```
cat /data/tools/setenv.sh >> ~/.bashrc
source ~/.bashrc
```





检查是否运行成功

```
conda info --envs #查看conda环境
ollama list # 查看ollama有哪些模型
ollama run bsahane/Qwen2.5-VL-7B-Instruct:Q4_K_M_benxh # 运行ollama交互式，ctrl d 取消
```

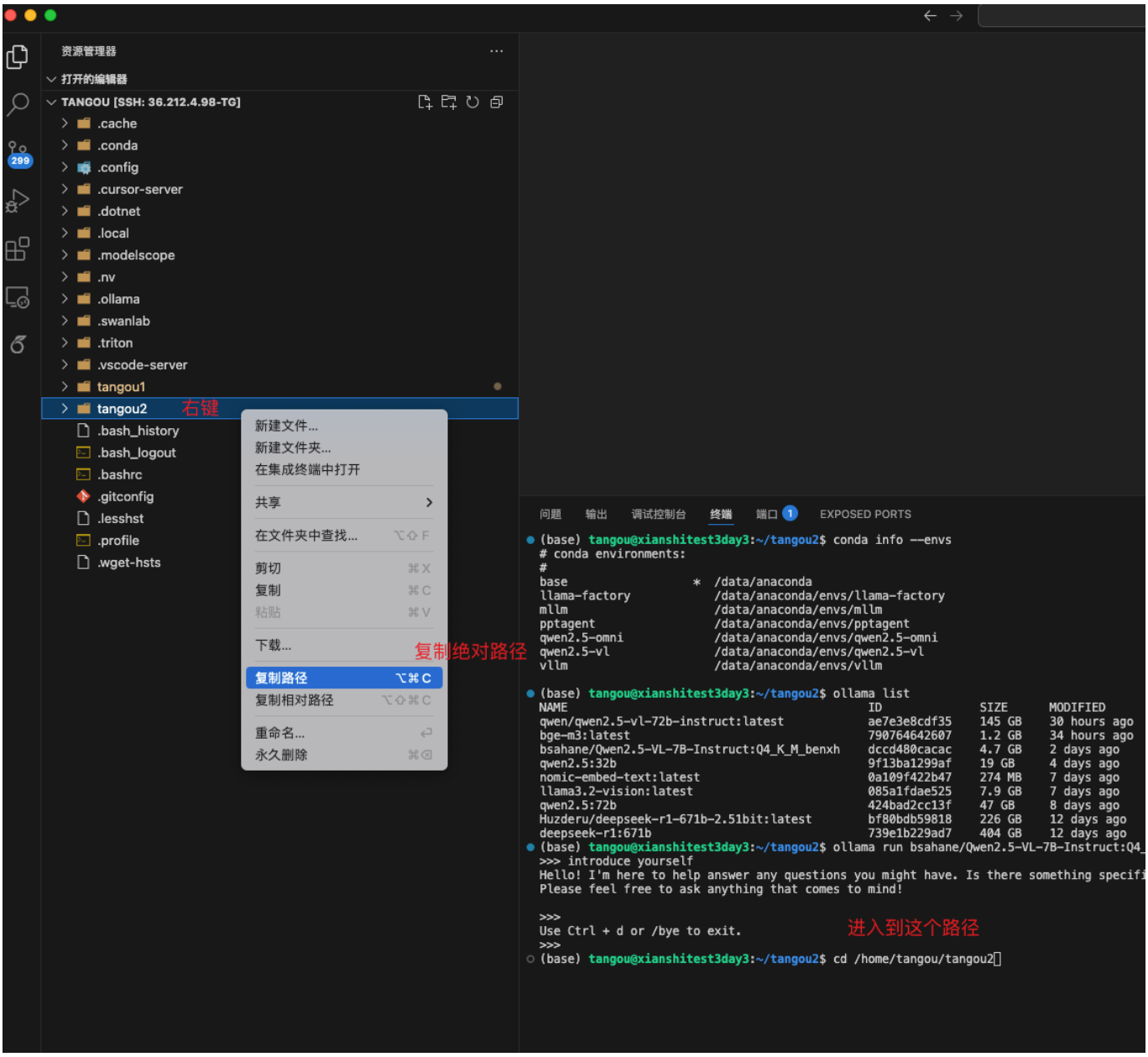
- vpn：先不用管，知道这个就行

```
# http://127.0.0.1:18099
source /data/tools/setproxy.sh  #启动vpn
source /data/tools/unsetproxy.sh  #关闭vpn
```

- python环境

1. 打开终端，进入文件夹

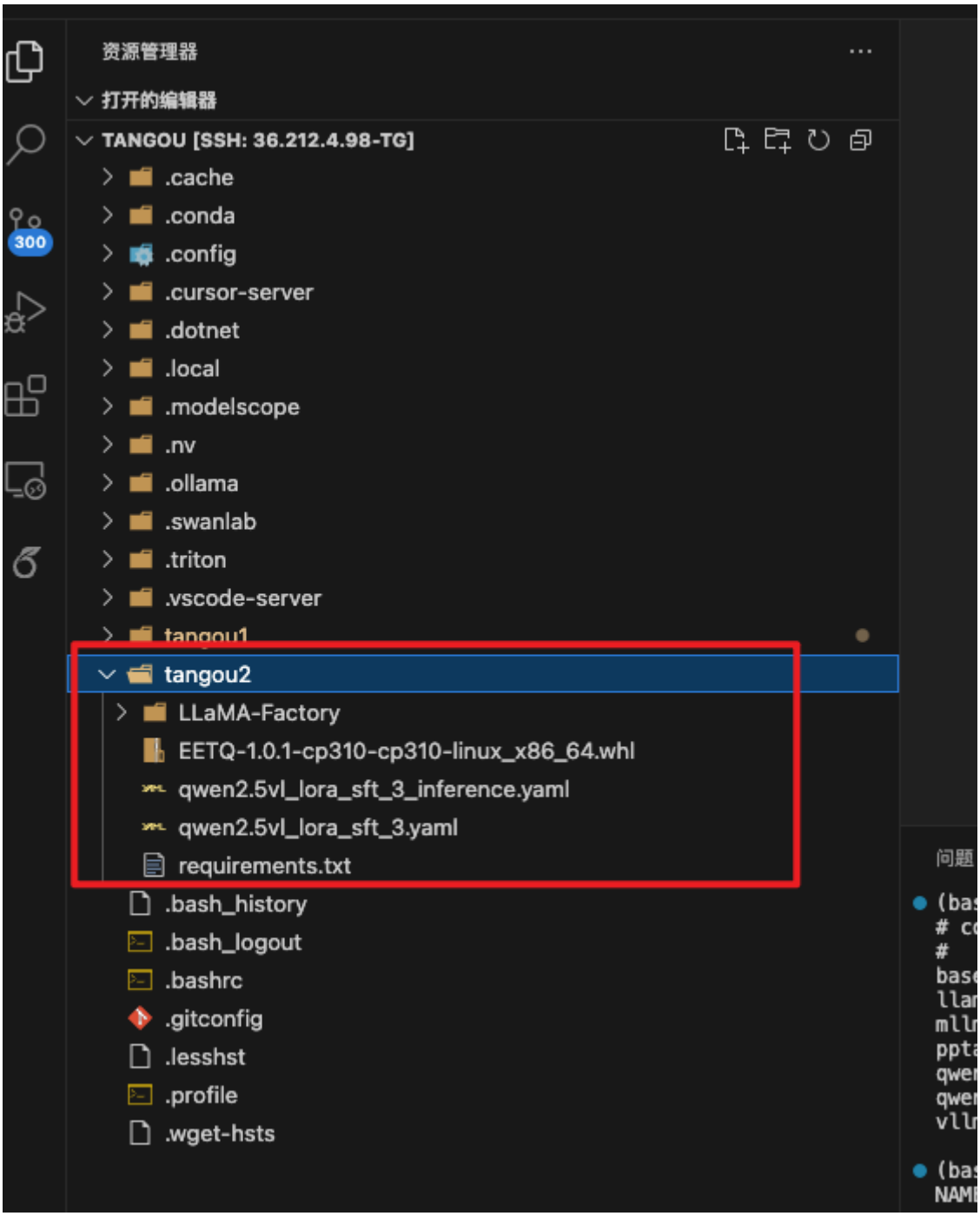
```
cd /home/tangou/tangou2 #你自己的路径
```



2. copy文件到目录

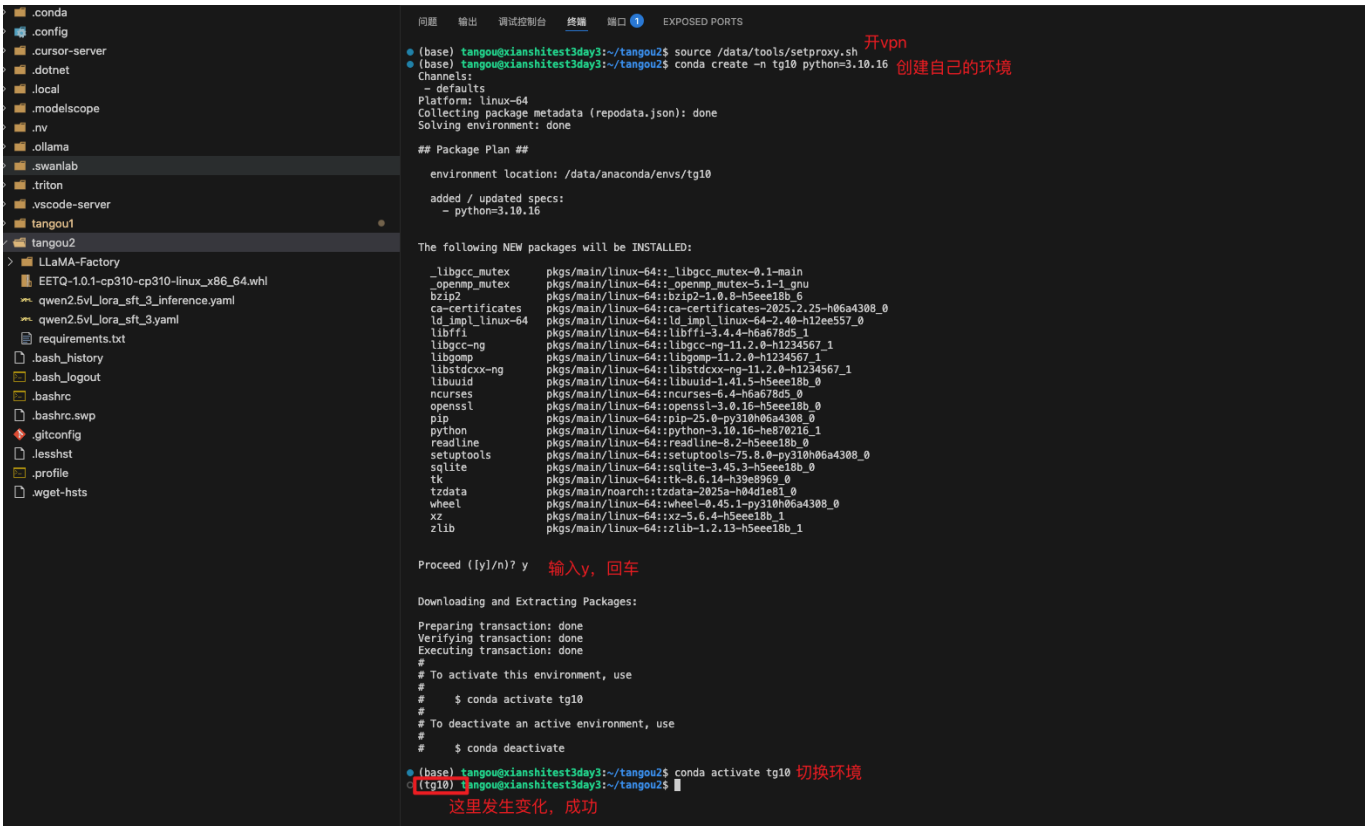
```
cp -r /data/3e/share/* /home/tangou/tangou2/
```

copy之后



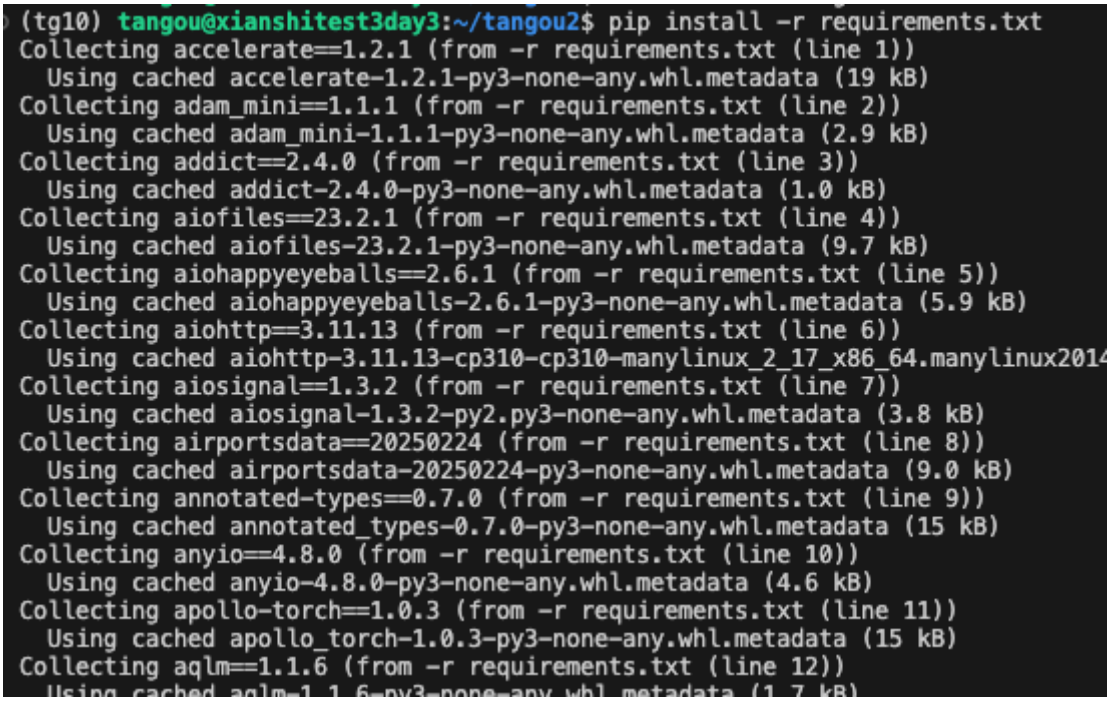
3. 创建环境并安装包

```
# 开vpn
source /data/tools/setproxy.sh
# tg10 换成自己的名字
conda create -n tg10 python=3.10.16
# 切换环境
conda activate tg10
```



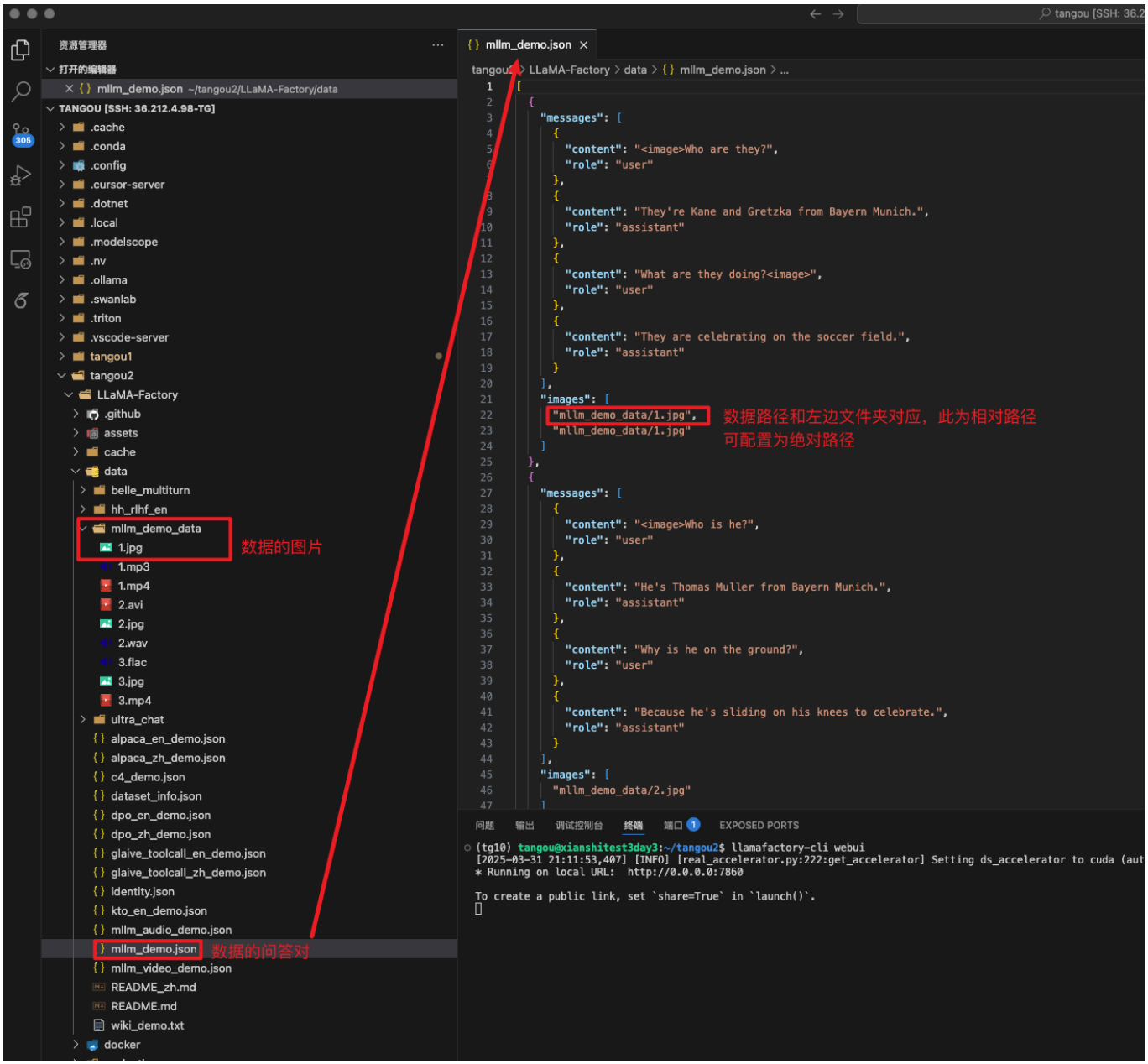
4. pip安装包

```
# 如果重新打开终端，没启动。请启动下，开vpn。
source /data/tools/setproxy.sh
# 安装包，第一次跑没缓存，运行时间会很久，在下数据包
pip install -r requirements.txt
# 额外安装这个包，pip源没有
pip install EETQ-1.0.1-cp310-cp310-linux_x86_64.whl
```

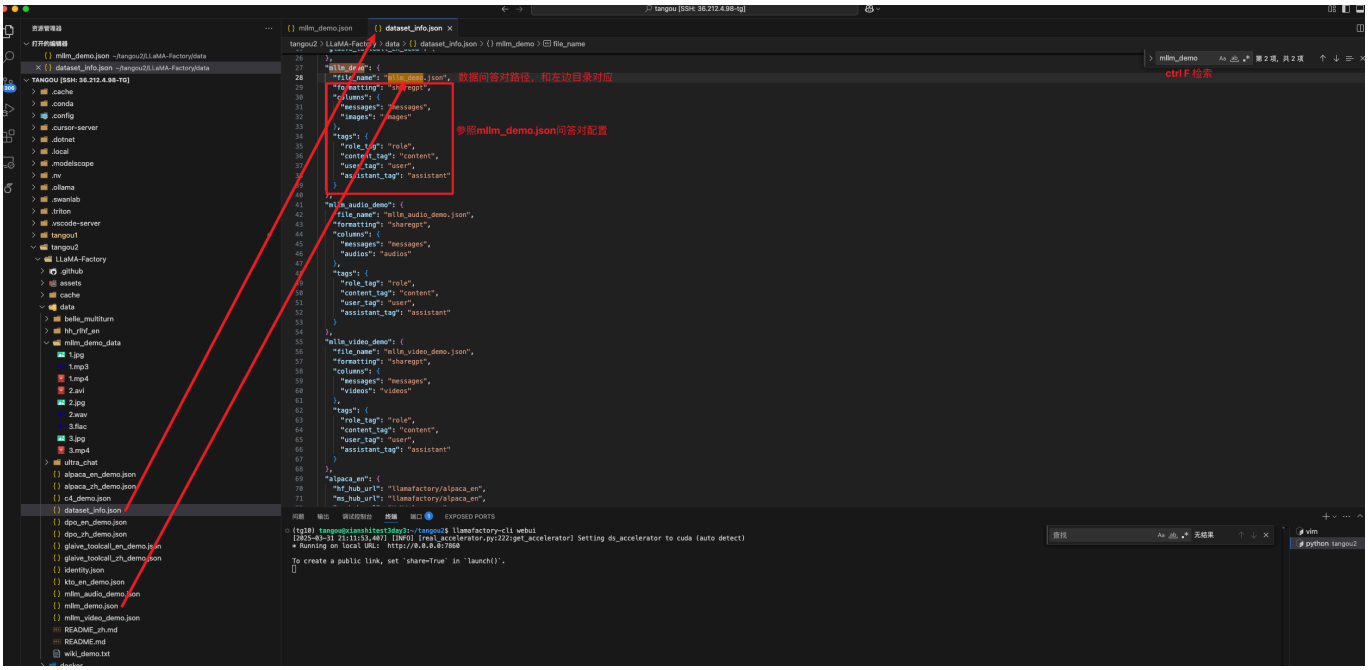


微调，这里用llamafactory提供的数据

1. 数据解读

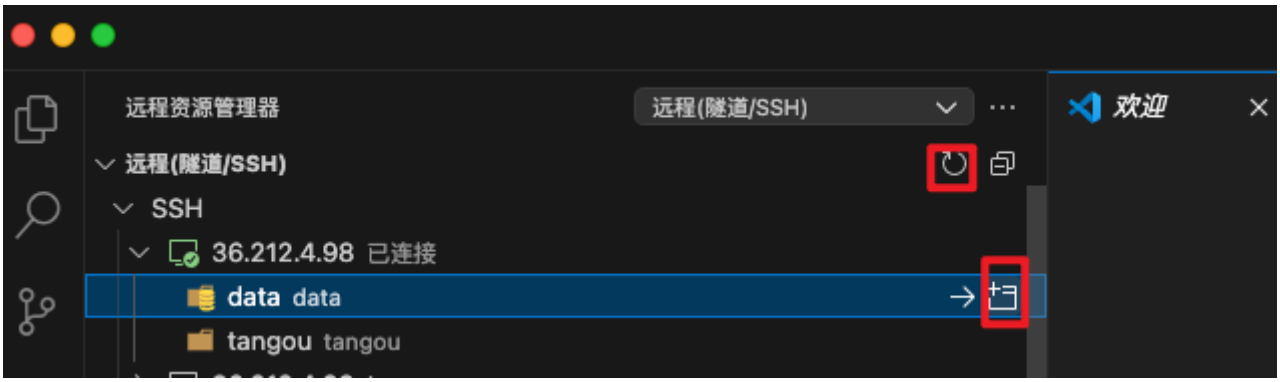


2. 微调加载数据，首先将自定义数据配置到dataset_info.json

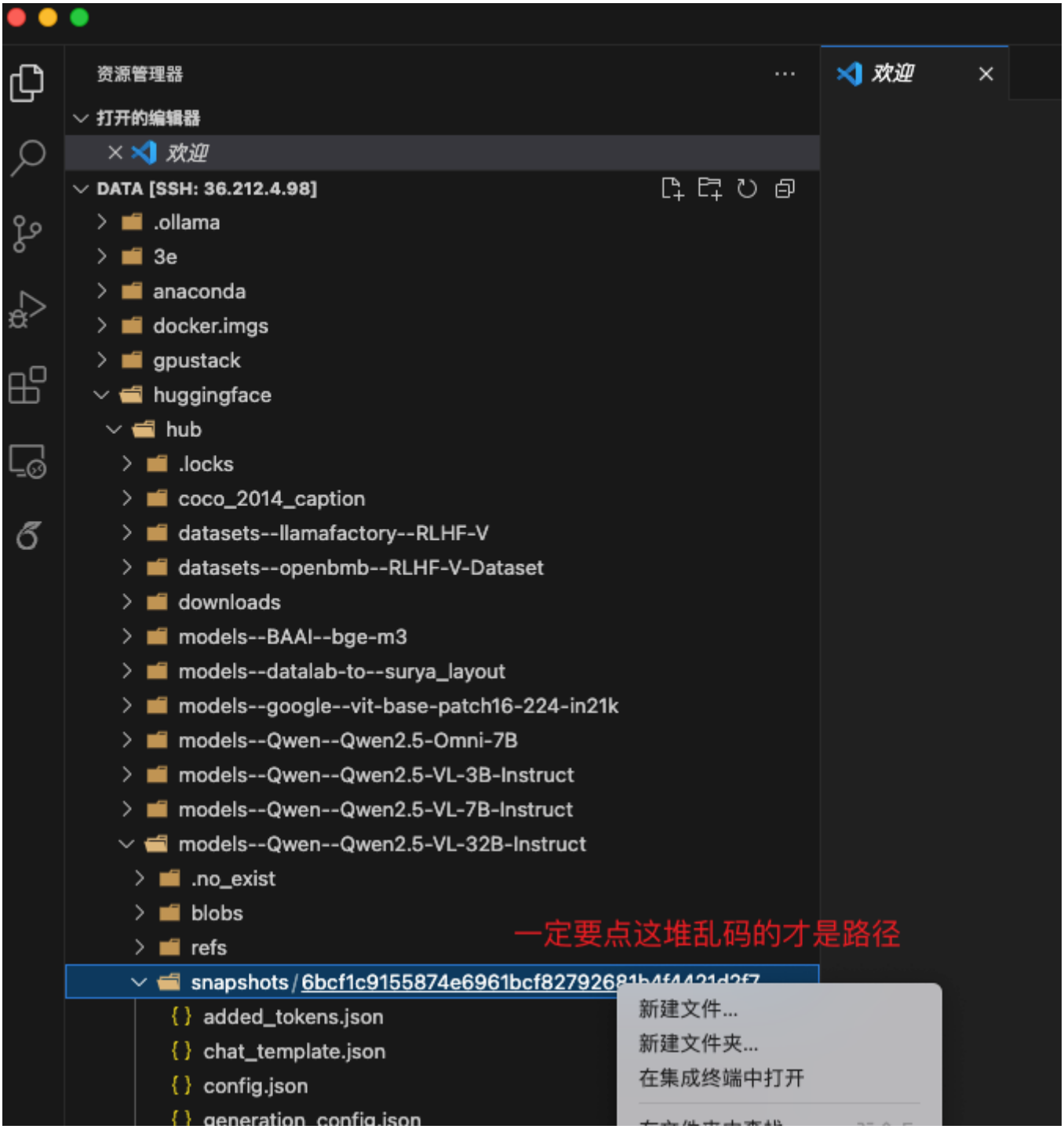


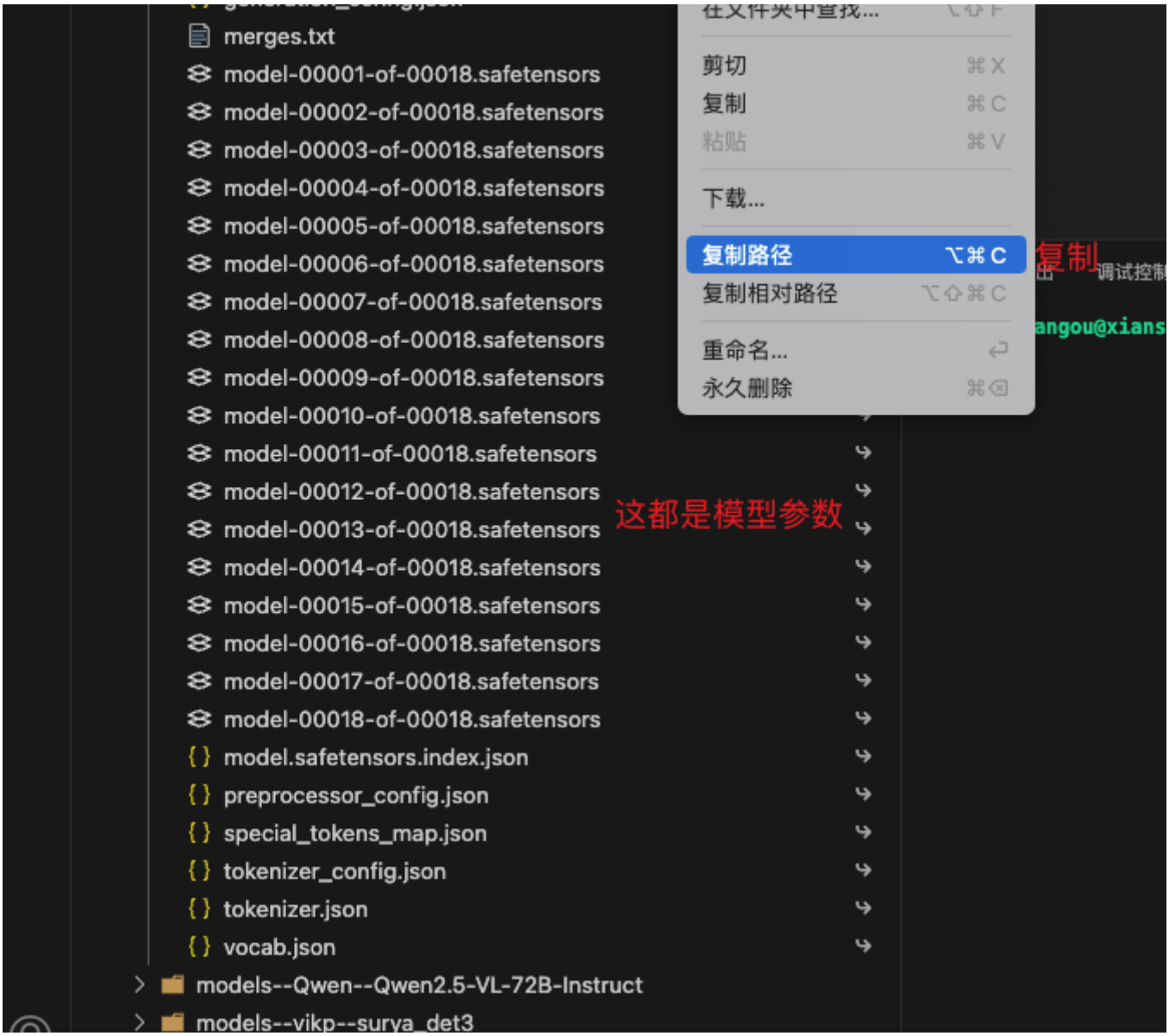
3. 配置模型路径

首先回到连接data共享目录，连vscode



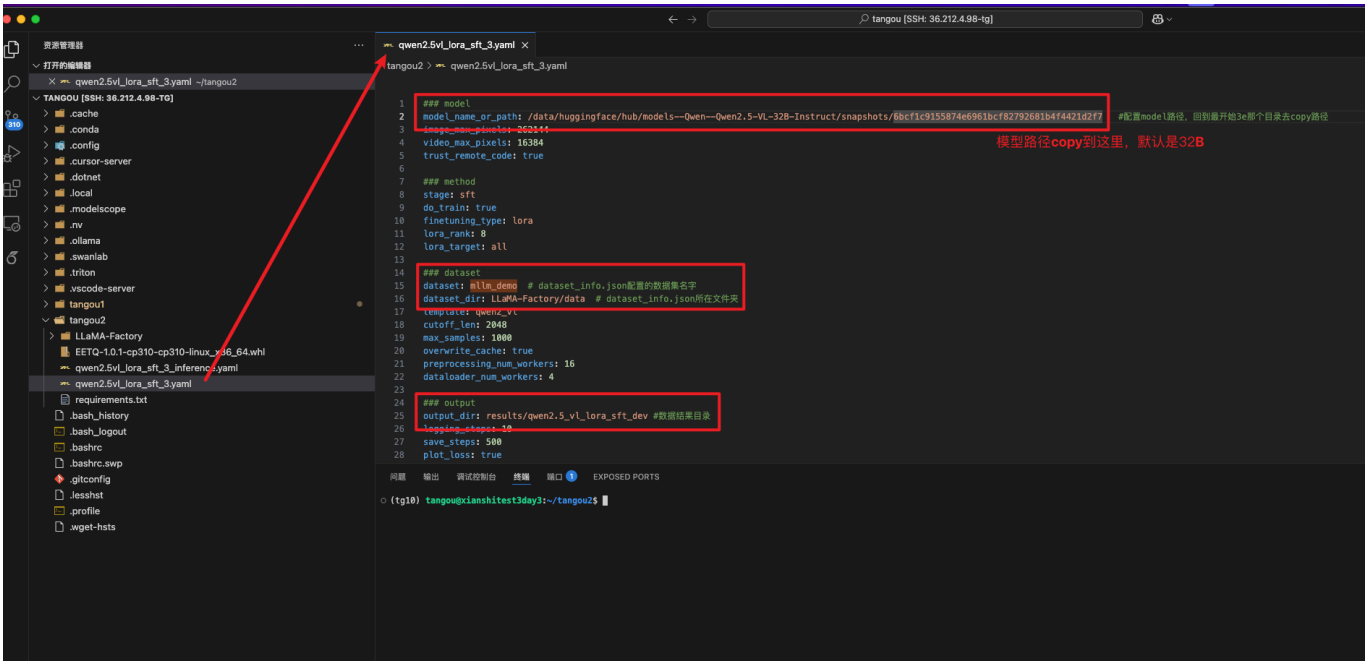
复制路径，我们这里微调32B





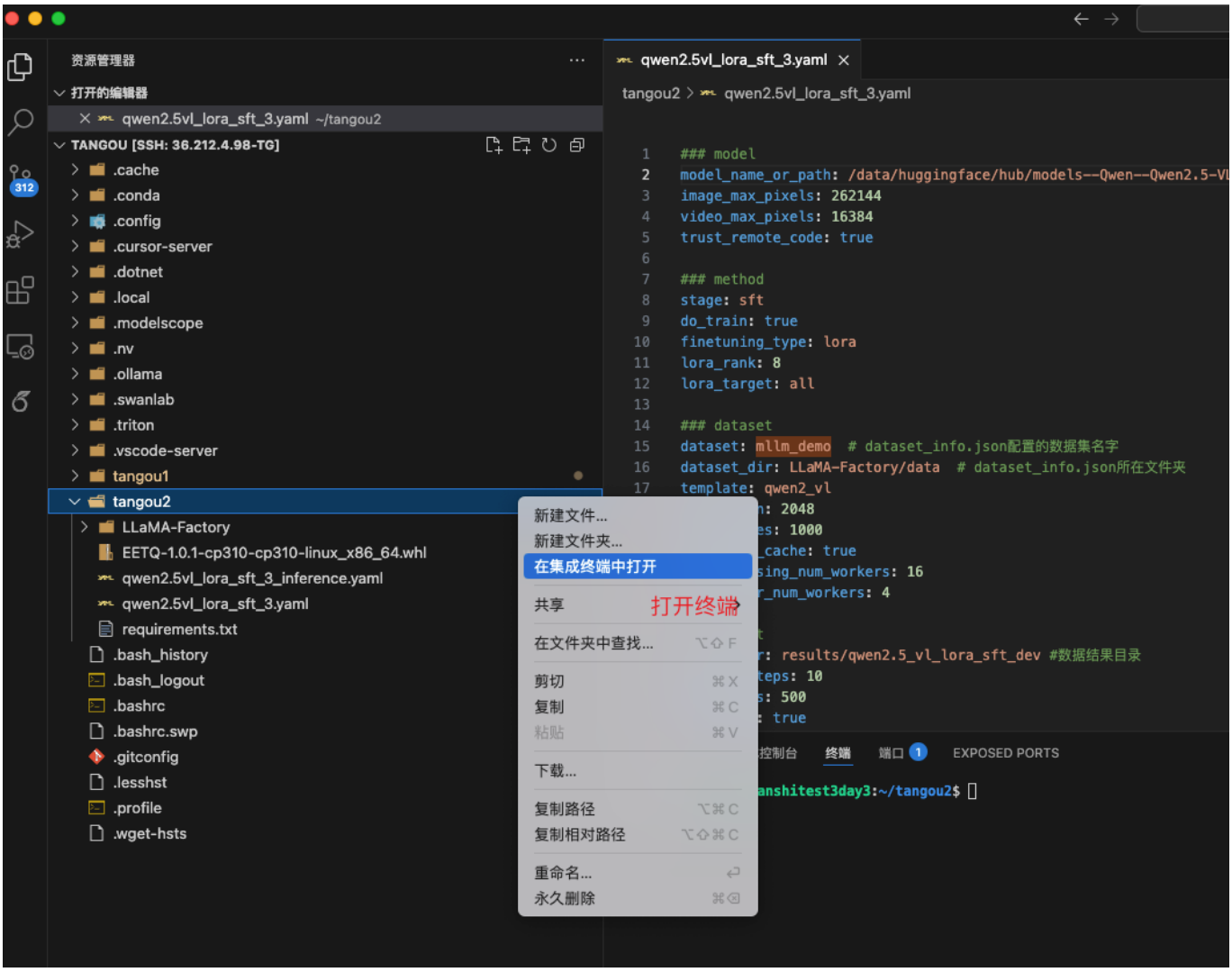
3. 配置微调的配置文件qwen2.5vl_lora_sft_3.yaml

回到原来的vscode，将上面复制的model路径放进来

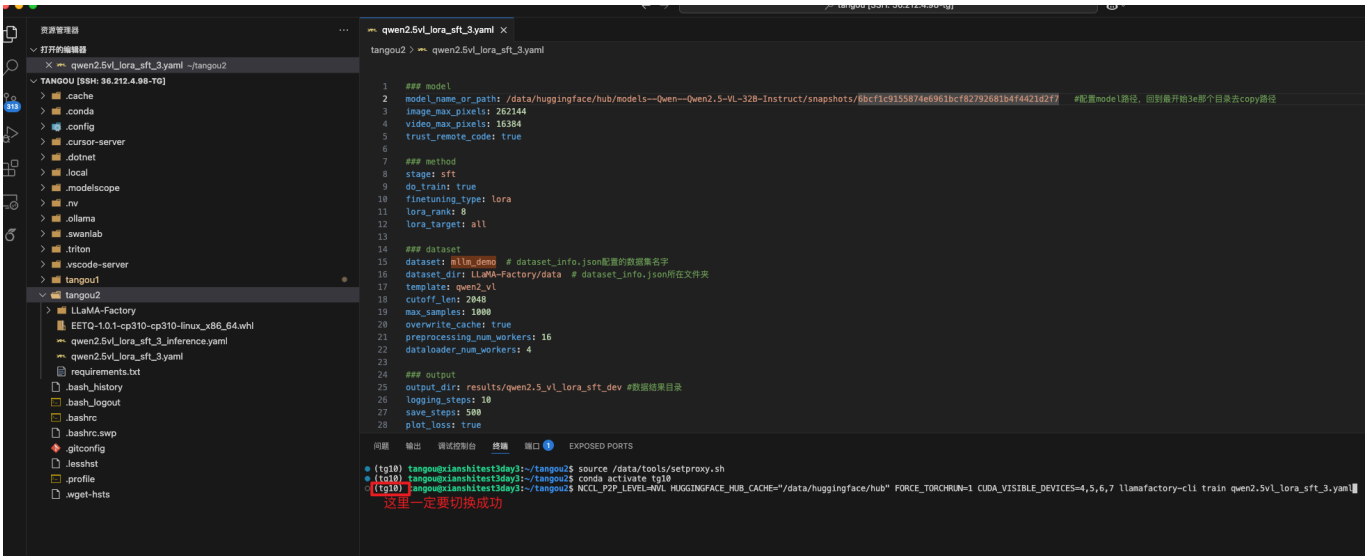


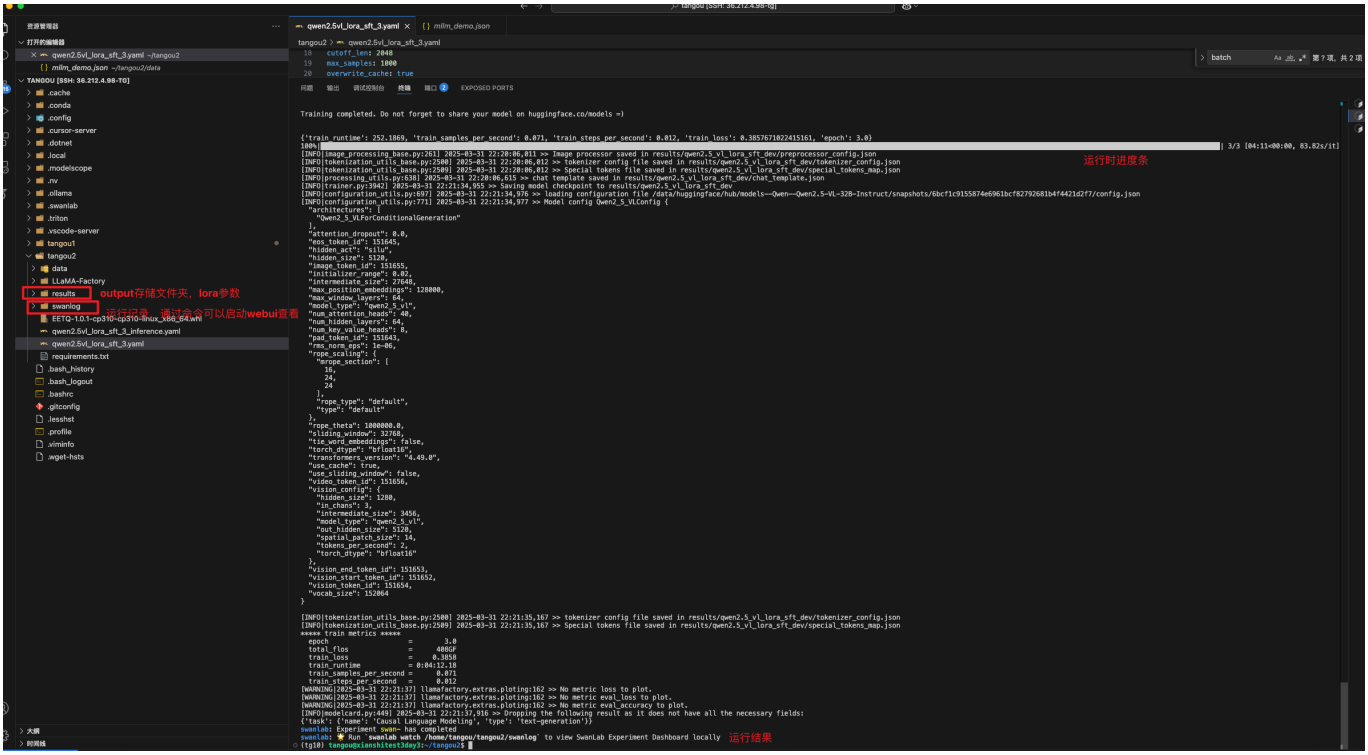
4. 运行微调

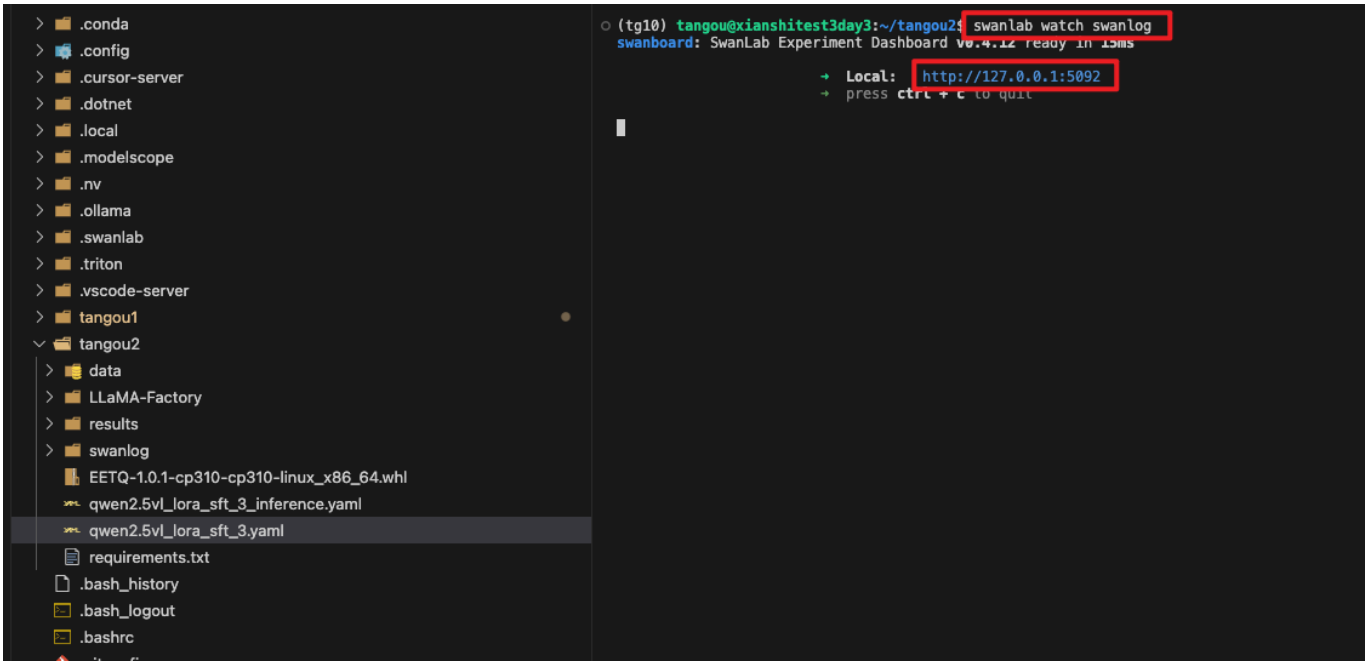
打开终端



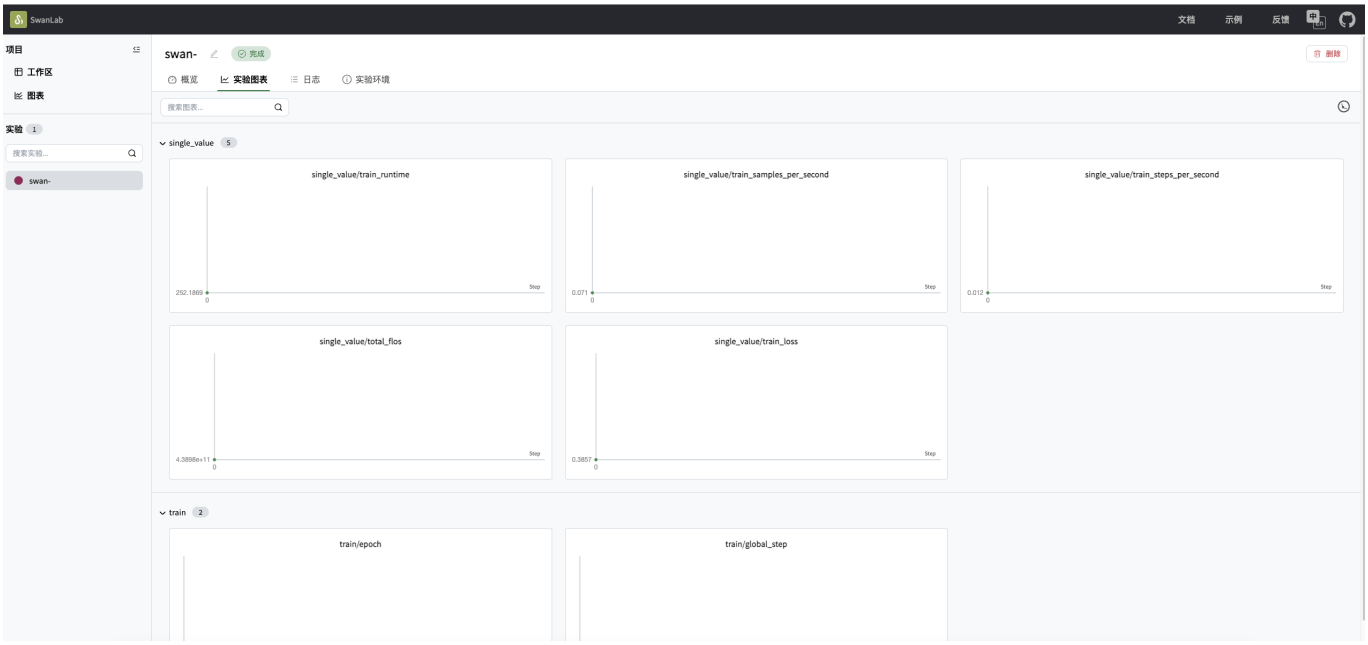

```
# 如果重新打开终端，没启动。请启动下，开vpn。
source /data/tools/setproxy.sh
# 切换你的python环境
conda activate tg10
# 训练
NCCL_P2P_LEVEL=NVL HUGGINGFACE_HUB_CACHE="/data/huggingface/hub"
FORCE_TORCHRUN=1 CUDA_VISIBLE_DEVICES=0,1,2,3,4,5,6,7 llamafactory-cli
train qwen2.5vl_lora_sft_3.yaml



# 查看运行记录，swanlog是相对路径，如果端口被占用，则--port xxx
conda activate tg10
swanlab watch swanlog --port 5092
```


这里一定要切换成功


output存储文件夹，lora参数
运行时进度条
运行结果

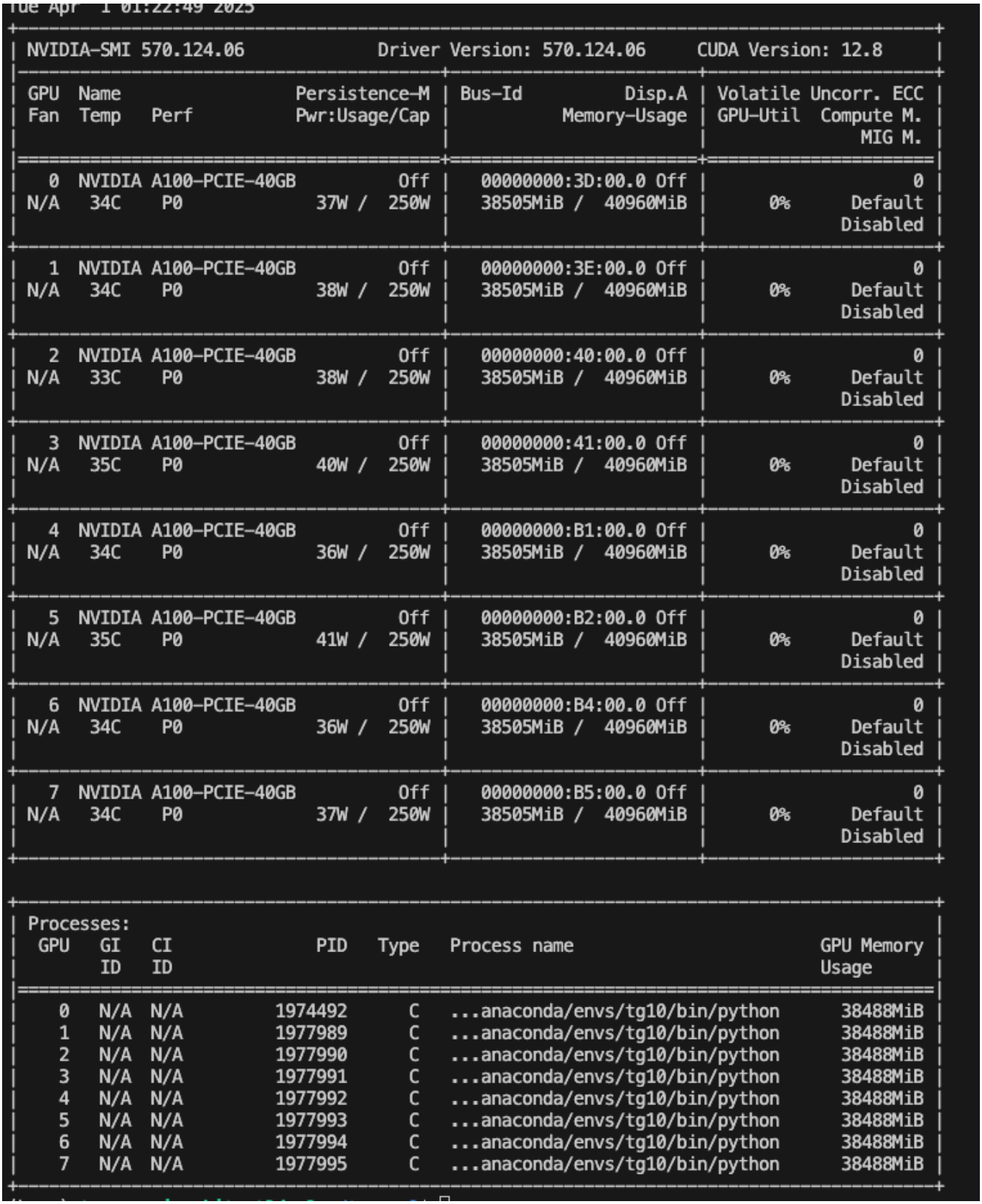

swanlab watch swanlog
Local: http://127.0.0.1:5092
press ctrl + c to quit

本地浏览器访问：http://127.0.0.1:5092



- 实测：32B
- 训练：在per_device_train_batch_size=1的情况下，32B显存空余如下，如果调72B需要乘2，72B勉强够用。根据显存空余，可调大per_device_train_batch_size=2, 4, 6, 8不等。
- 训练：6组数据（每组2-3轮对话），一个epoch需要：50s-80s。
- 推理：差点爆显存
- 评估：直接爆显存

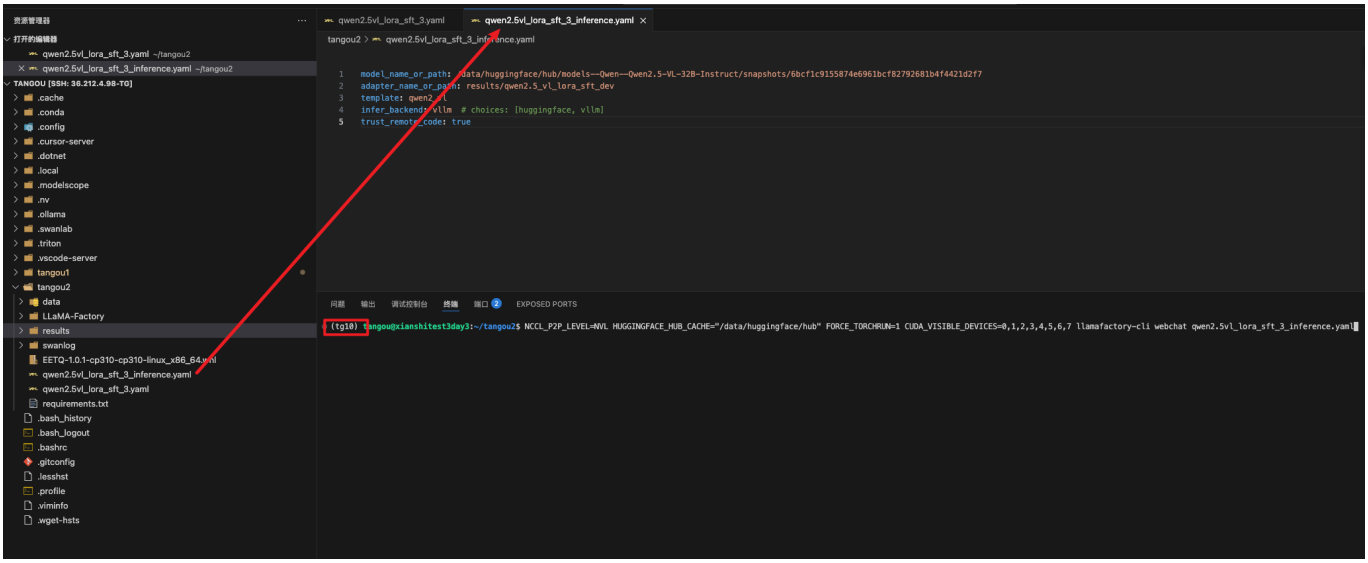
```
Mon Mar 31 22:18:35 2025
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 570.124.06             Driver Version: 570.124.06     CUDA Version: 12.8     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA A100-PCIE-40GB          Off |   00000000:3D:00.0 Off |                    0 |
| N/A   38C    P0             70W /  250W |   17887MiB /  40960MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   1  NVIDIA A100-PCIE-40GB          Off |   00000000:3E:00.0 Off |                    0 |
| N/A   38C    P0             74W /  250W |   17883MiB /  40960MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   2  NVIDIA A100-PCIE-40GB          Off |   00000000:40:00.0 Off |                    0 |
| N/A   37C    P0             70W /  250W |   18033MiB /  40960MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   3  NVIDIA A100-PCIE-40GB          Off |   00000000:41:00.0 Off |                    0 |
| N/A   39C    P0             88W /  250W |   17947MiB /  40960MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   4  NVIDIA A100-PCIE-40GB          Off |   00000000:B1:00.0 Off |                    0 |
| N/A   37C    P0             68W /  250W |   17933MiB /  40960MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   5  NVIDIA A100-PCIE-40GB          Off |   00000000:B2:00.0 Off |                    0 |
| N/A   39C    P0             99W /  250W |   16801MiB /  40960MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   6  NVIDIA A100-PCIE-40GB          Off |   00000000:B4:00.0 Off |                    0 |
| N/A   38C    P0             75W /  250W |   17887MiB /  40960MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+
|   7  NVIDIA A100-PCIE-40GB          Off |   00000000:B5:00.0 Off |                    0 |
| N/A   38C    P0             70W /  250W |   17885MiB /  40960MiB |    100%      Default |
|                                         |                        |             Disabled |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI              PID   Type   Process name                        GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|    0   N/A  N/A          382847      C   ...anaconda/envs/tg10/bin/python      17878MiB |
|    1   N/A  N/A          382848      C   ...anaconda/envs/tg10/bin/python      17874MiB |
|    2   N/A  N/A          382849      C   ...anaconda/envs/tg10/bin/python      18024MiB |
|    3   N/A  N/A          382850      C   ...anaconda/envs/tg10/bin/python      17938MiB |
|    4   N/A  N/A          382851      C   ...anaconda/envs/tg10/bin/python      17924MiB |
|    5   N/A  N/A          382852      C   ...anaconda/envs/tg10/bin/python      16792MiB |
|    6   N/A  N/A          382853      C   ...anaconda/envs/tg10/bin/python      17878MiB |
|    7   N/A  N/A          382854      C   ...anaconda/envs/tg10/bin/python      17876MiB |
+-----------------------------------------------------------------------------------------+
```

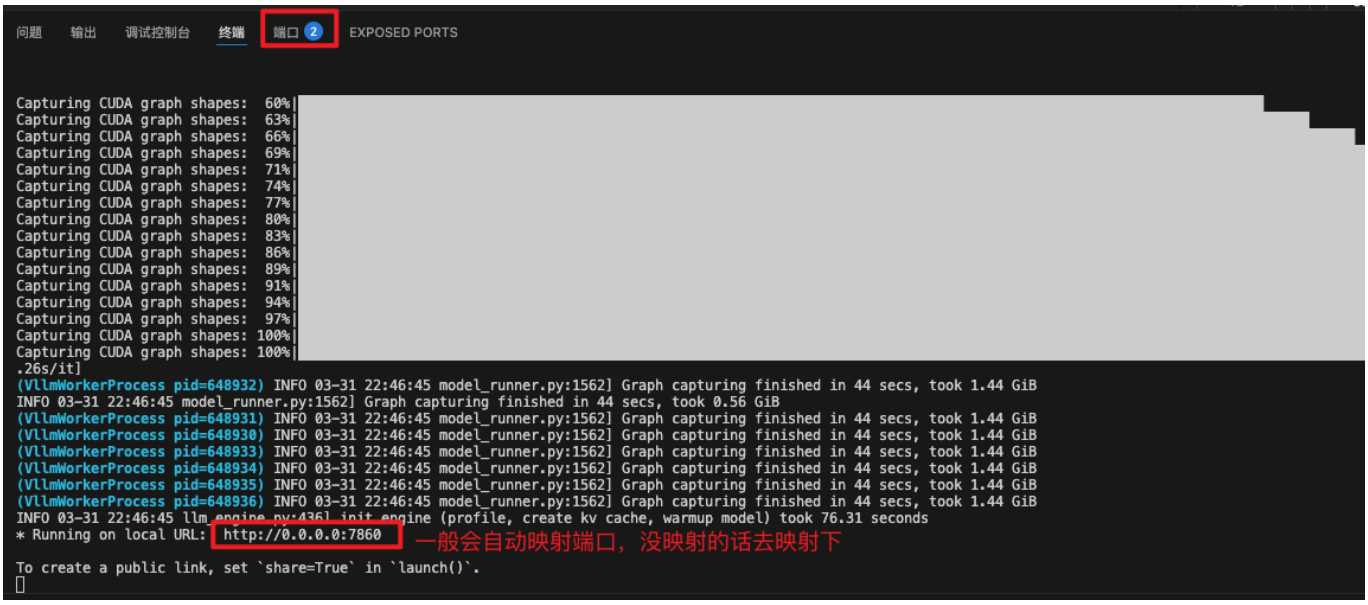


5. 推理

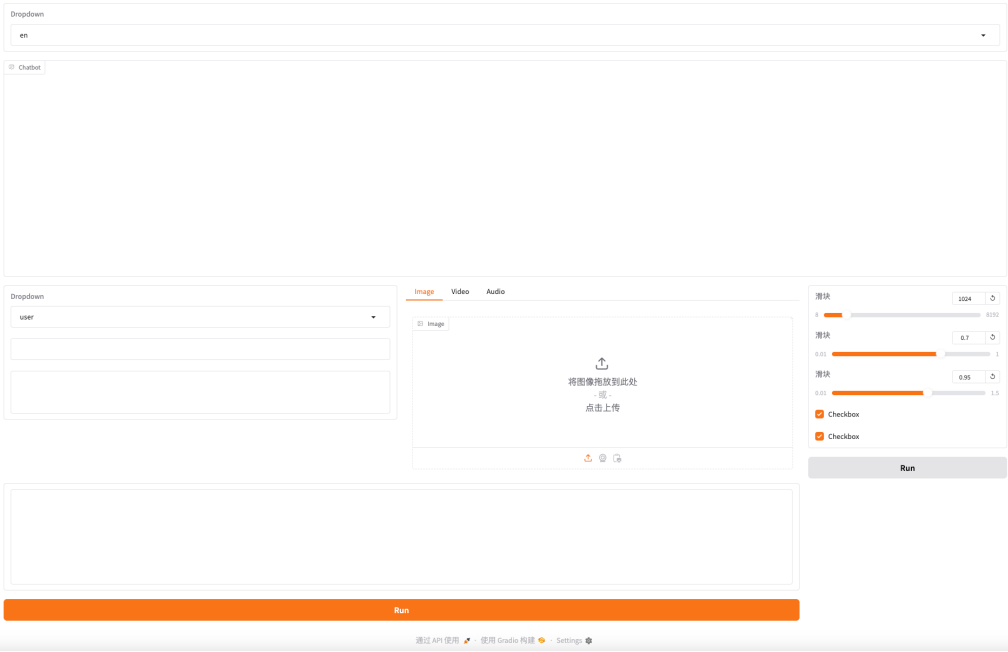
```
source /data/tools/setproxy.sh
conda activate tg10
# 如果端口占用，请换个端口
export GRADIO_SERVER_PORT=7860
NCCL_P2P_LEVEL=NVL HUGGINGFACE_HUB_CACHE="/data/huggingface/hub"
```

```
FORCE_TORCHRUN=1 CUDA_VISIBLE_DEVICES=0,1,2,3,4,5,6,7 llamafactory-cli
webchat qwen2.5vl_lora_sft_3_inference.yaml
```

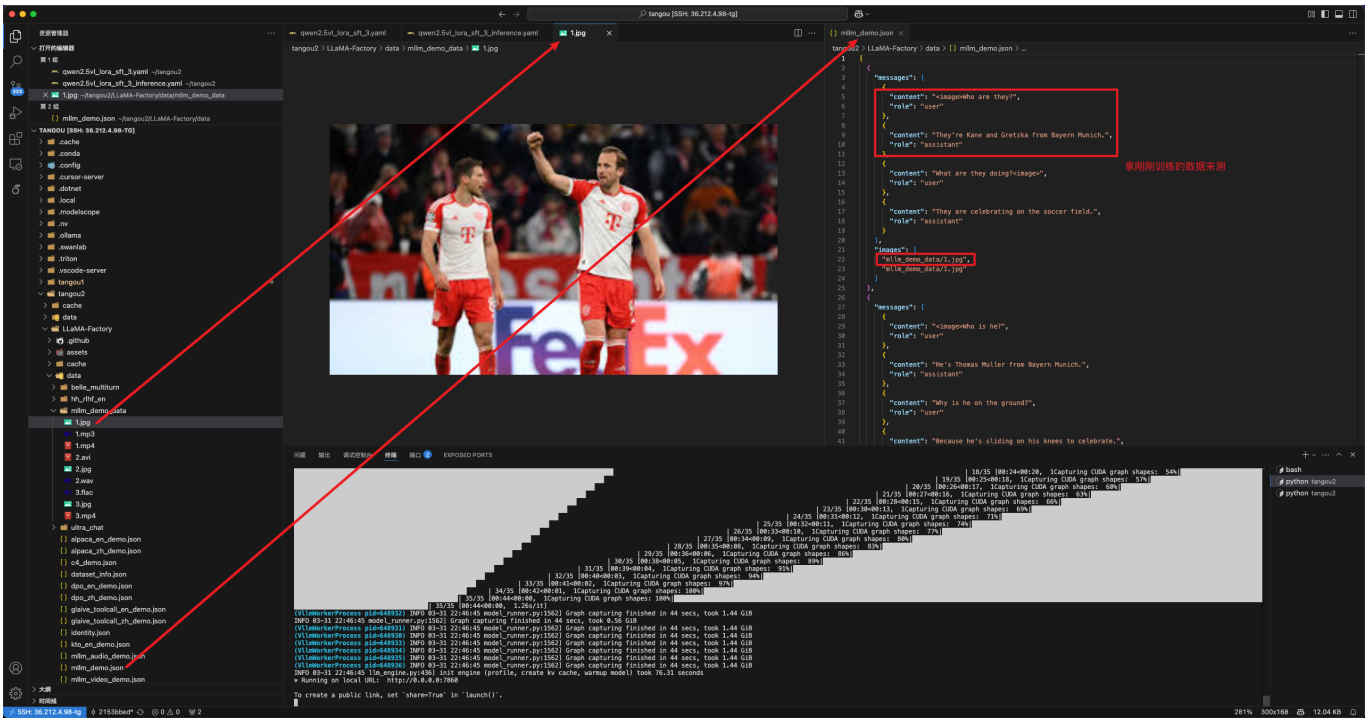




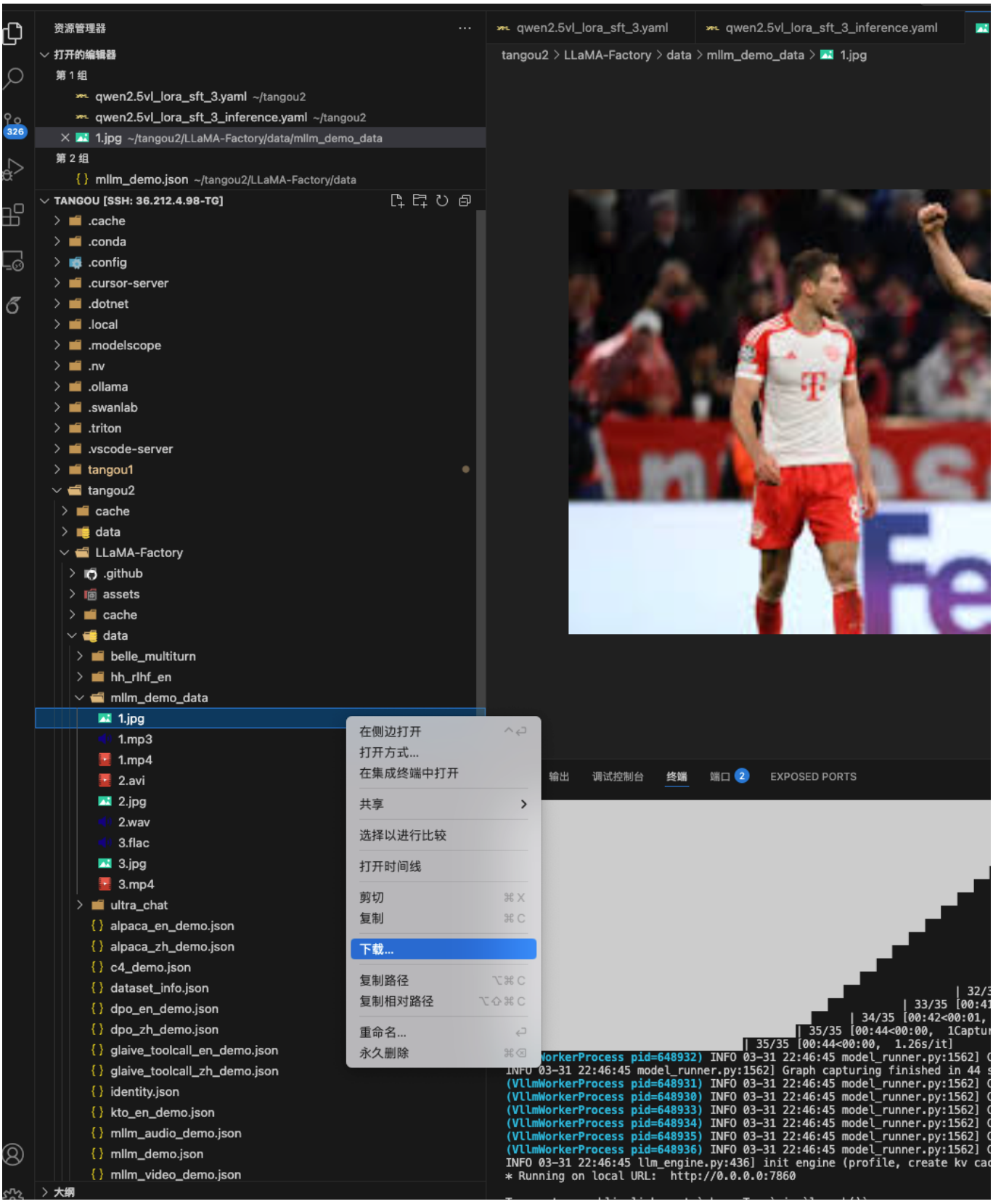
本地浏览器访问：http://0.0.0.0:7860



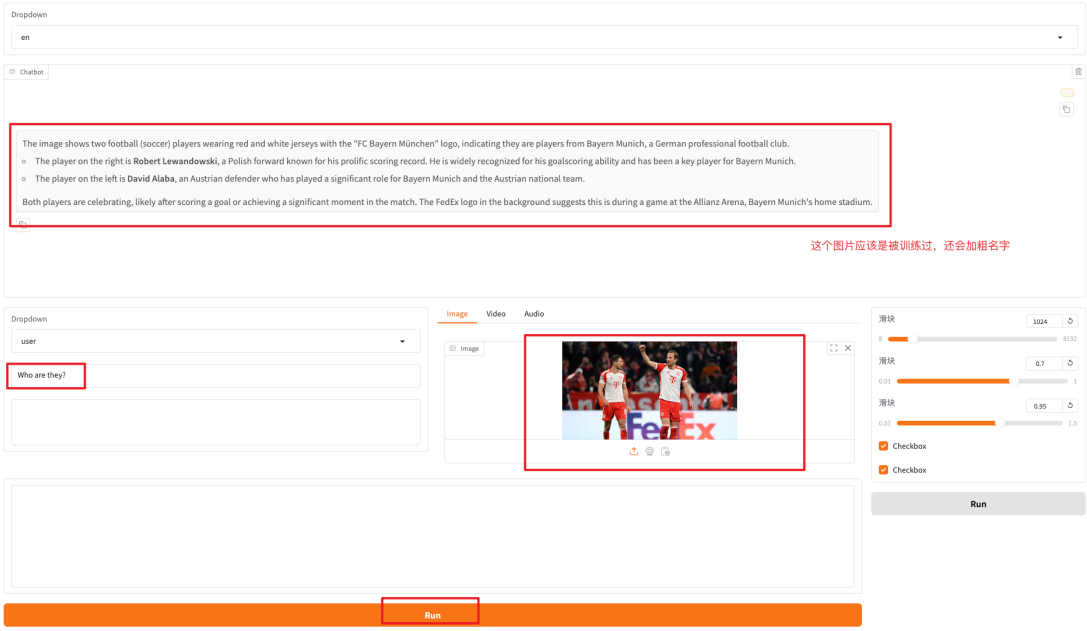
拿刚刚训练的数据来测



下载到本地

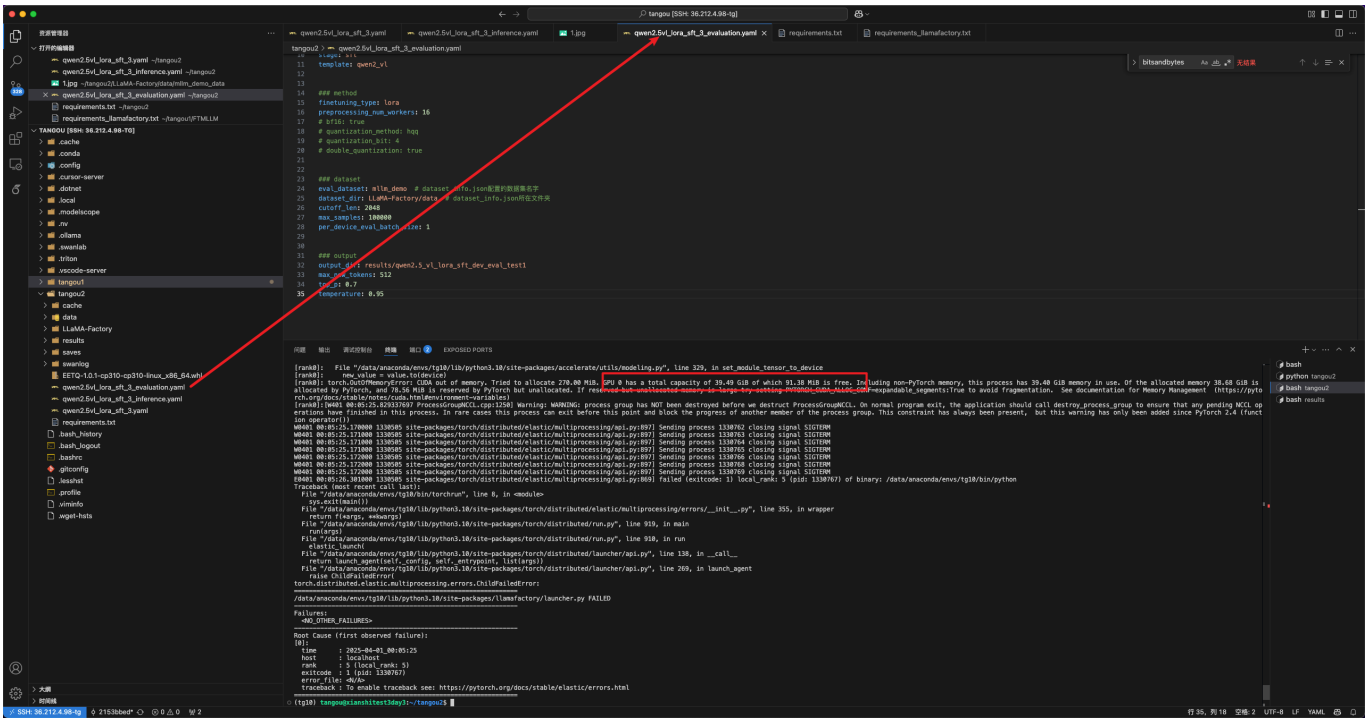


推理（这个图片在原本的模型上就一个训练过）



6. 评估（实测：32B评估时会爆显存）

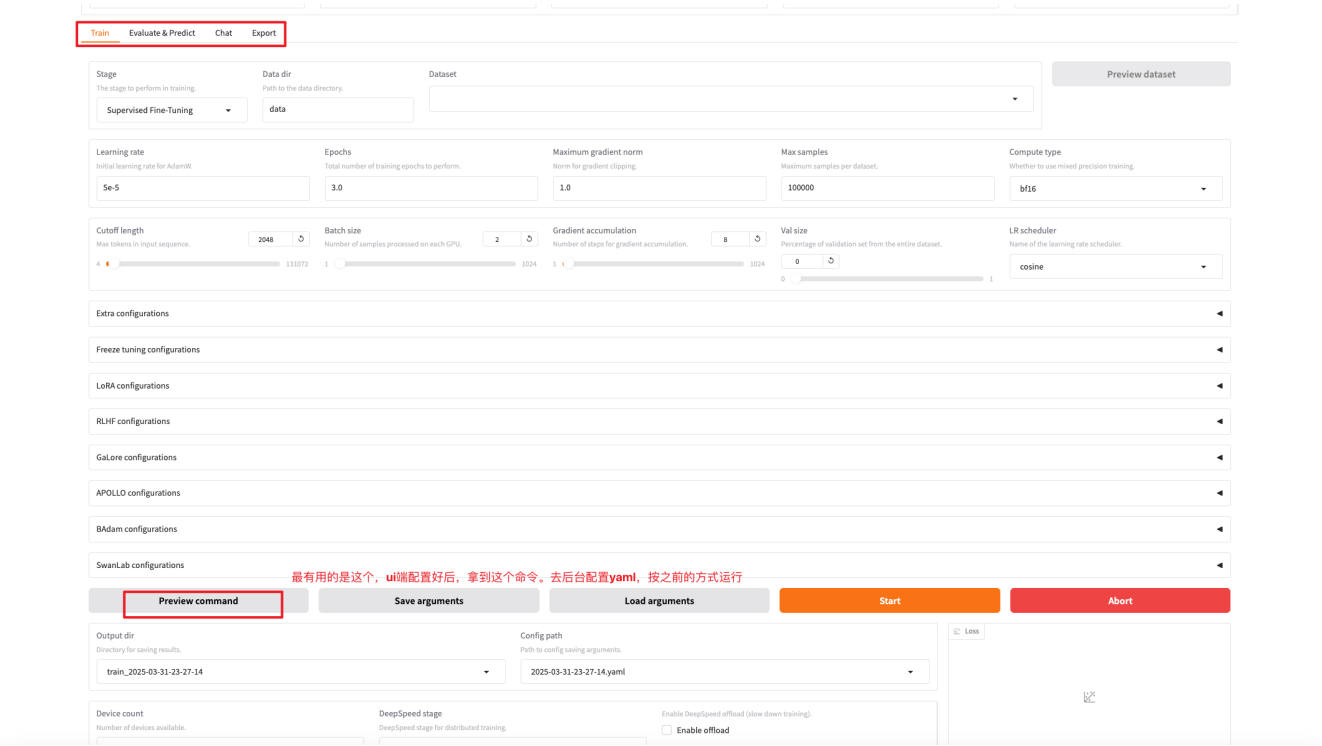
```
source /data/tools/setproxy.sh
conda activate tg10
NCCL_P2P_LEVEL=NVL HUGGINGFACE_HUB_CACHE="/data/huggingface/hub"
FORCE_TORCHRUN=1 CUDA_VISIBLE_DEVICES=0,1,2,3,4,5,6,7 llamafactory-cli
train qwen2.5vl_lora_sft_3_evaluation.yaml
```



额外

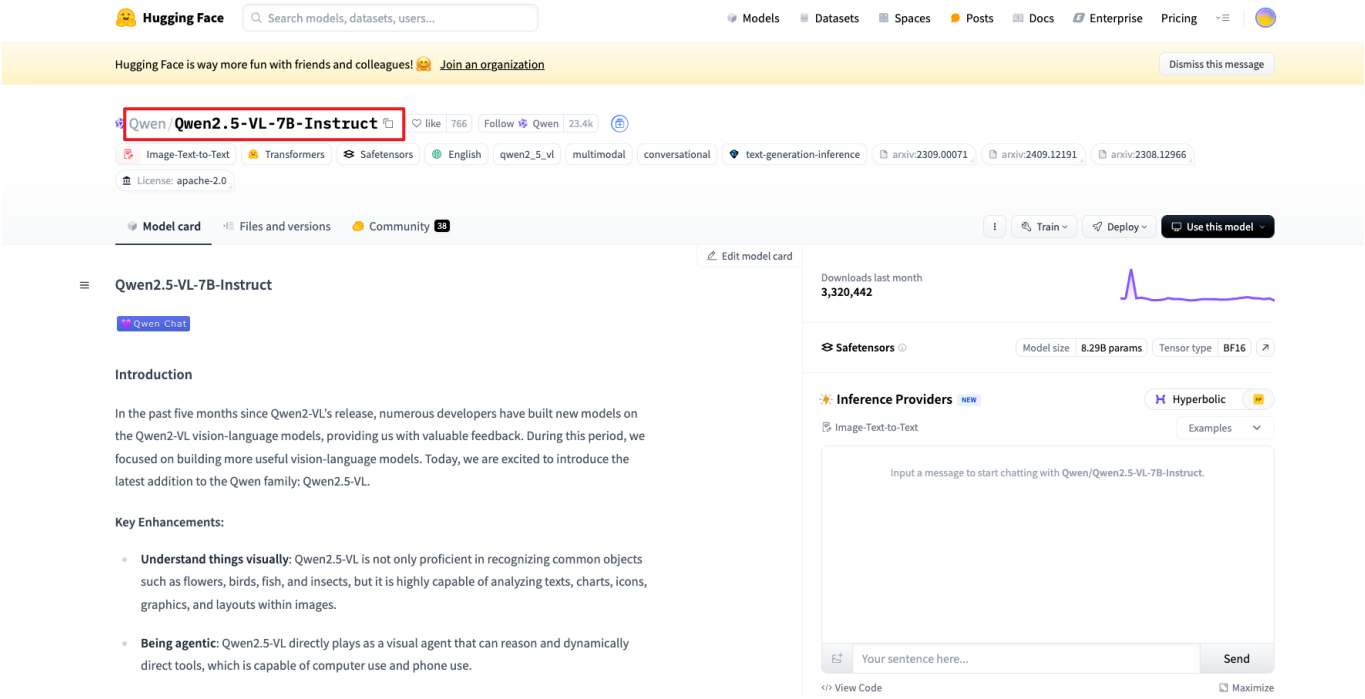
webui运行

```
source /data/tools/setproxy.sh
conda activate tg10
# 如果端口占用请换个端口
export GRADIO_SERVER_PORT=7860
llamafactory-cli webui
```



下载模型、数据集

```
source /data/tools/setproxy.sh
conda activate tg10
huggingface-cli login # token教程:
https://blog.csdn.net/m0_52625549/article/details/134255660
----
export HUGGINGFACE_HUB_CACHE="/data/huggingface/hub"  #设置缓存路径，就是之前的
共享目录
# 数据集
huggingface-cli download --resume-download --repo-type dataset
llamafactory/RLHF-V --local-dir-use-symlinks False
# 模型
huggingface-cli download --resume-download Qwen/Qwen2.5-VL-7B-Instruct --
local-dir-use-symlinks False
```


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Enterprise
Pricing
Hugging Face is way more fun with friends and colleagues! Join an organization
Dismiss this message
Qwen / Qwen2.5-VL-7B-Instruct
like 766
Follow Qwen 23.4k
Image-Text-to-Text
Transformers
Safetensors
English
qwen2_5_vl
multimodal
conversational
text-generation-inference
arxiv:2309.00071
arxiv:2409.12191
arxiv:2308.12966
License: apache-2.0
Model card
Files and versions
Community 38
Train
Deploy
Use this model
Edit model card
Qwen2.5-VL-7B-Instruct
Qwen Chat
Introduction
In the past five months since Qwen2-VL's release, numerous developers have built new models on the Qwen2-VL vision-language models, providing us with valuable feedback. During this period, we focused on building more useful vision-language models. Today, we are excited to introduce the latest addition to the Qwen family: Qwen2.5-VL.
Key Enhancements:
Understand things visually: Qwen2.5-VL is not only proficient in recognizing common objects such as flowers, birds, fish, and insects, but it is highly capable of analyzing texts, charts, icons, graphics, and layouts within images.
Being agentic: Qwen2.5-VL directly plays as a visual agent that can reason and dynamically direct tools, which is capable of computer use and phone use.
Downloads last month
3,320,442
Safetensors
Model size 8.29B params
Tensor type BF16
Inference Providers NEW
Hyperbolic
Image-Text-to-Text
Examples
Input a message to start chatting with Qwen/Qwen2.5-VL-7B-Instruct.
Your sentence here...
Send
View Code
Maximize