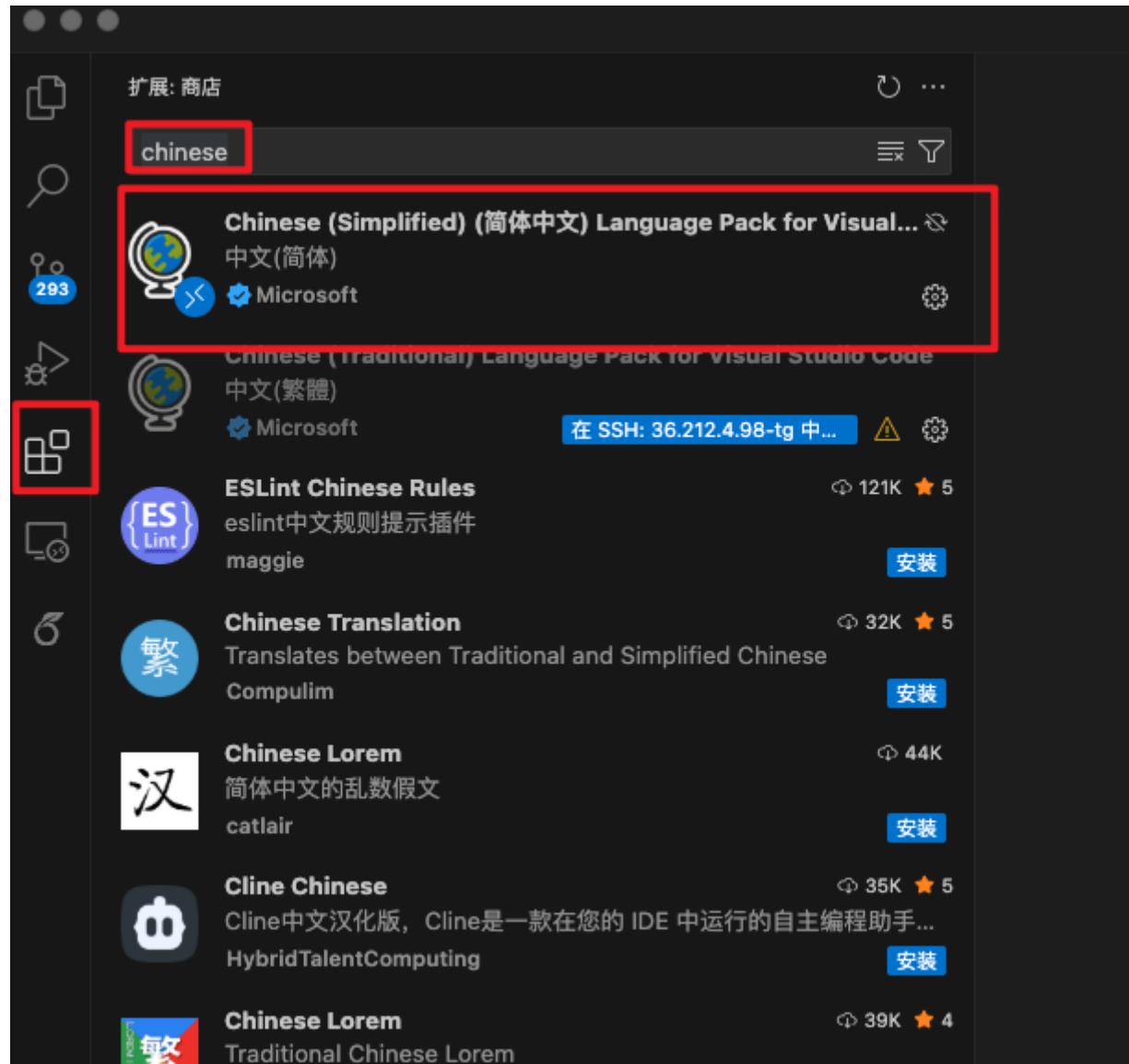


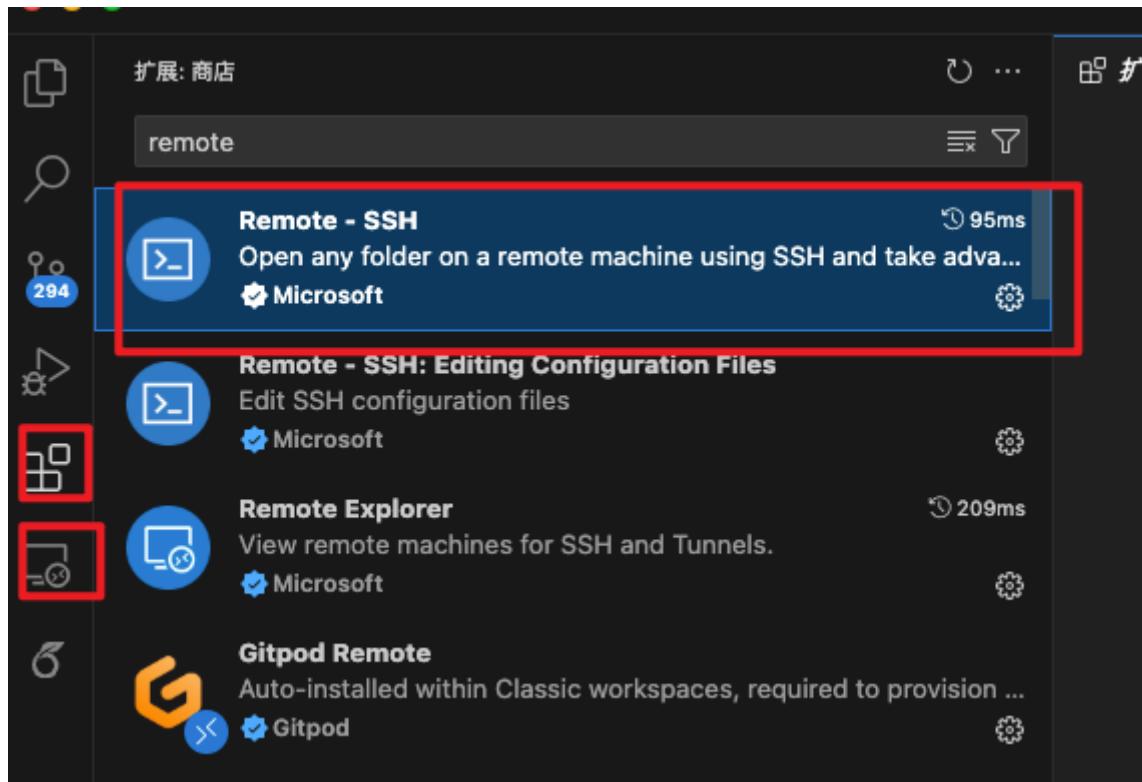
教程详细版

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

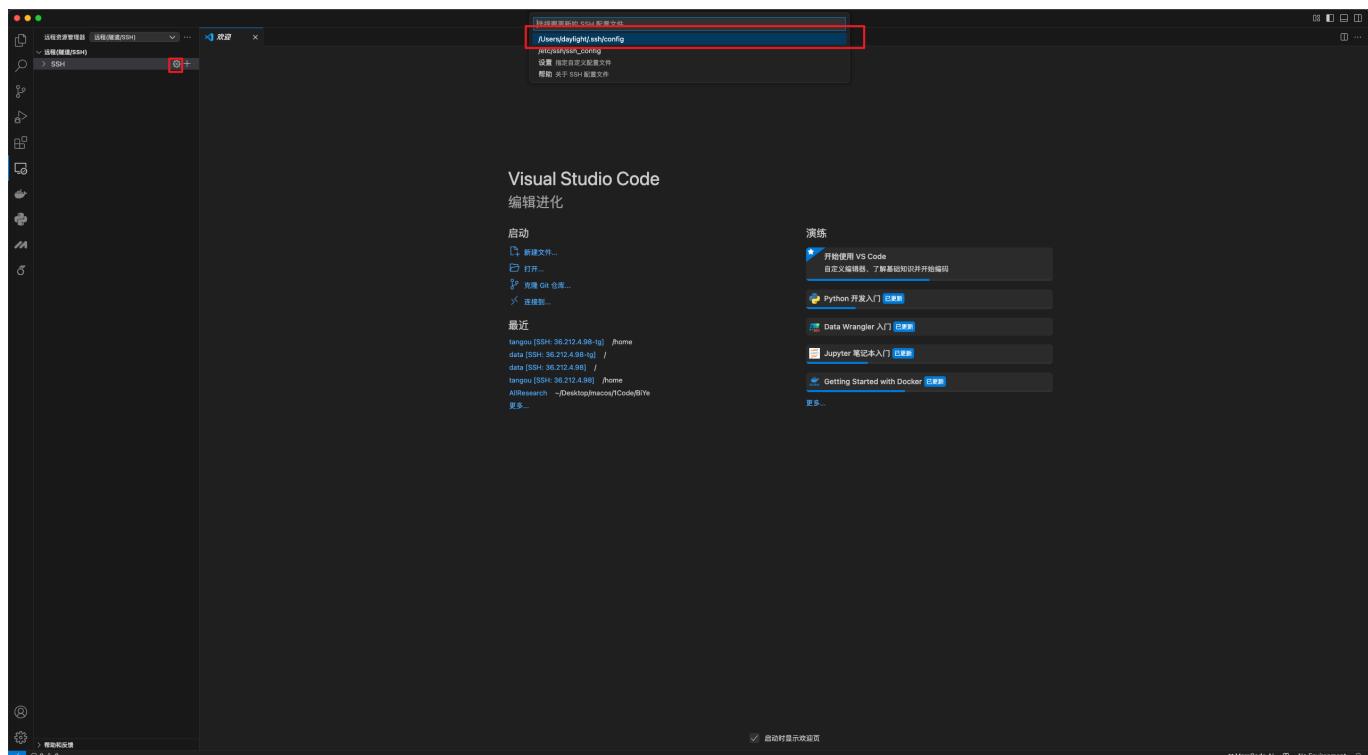




服务器连接

大家查看群文件自己的user和密码

vscode连接:



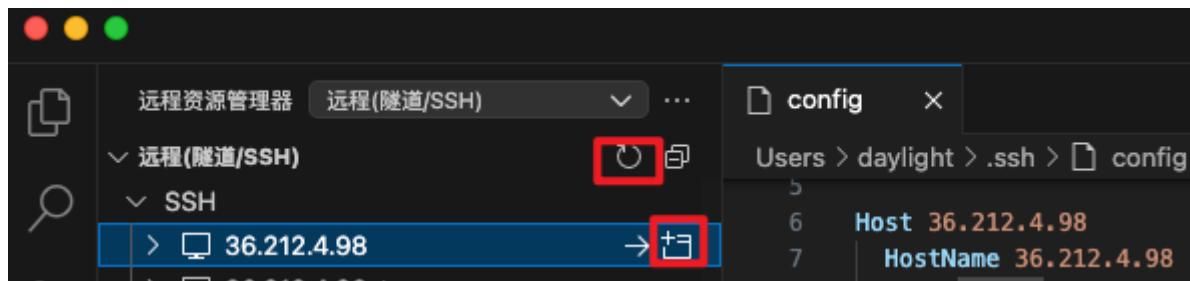
1. 复制到文件里面去

```
Host 36.212.4.98
HostName 36.212.4.98
```

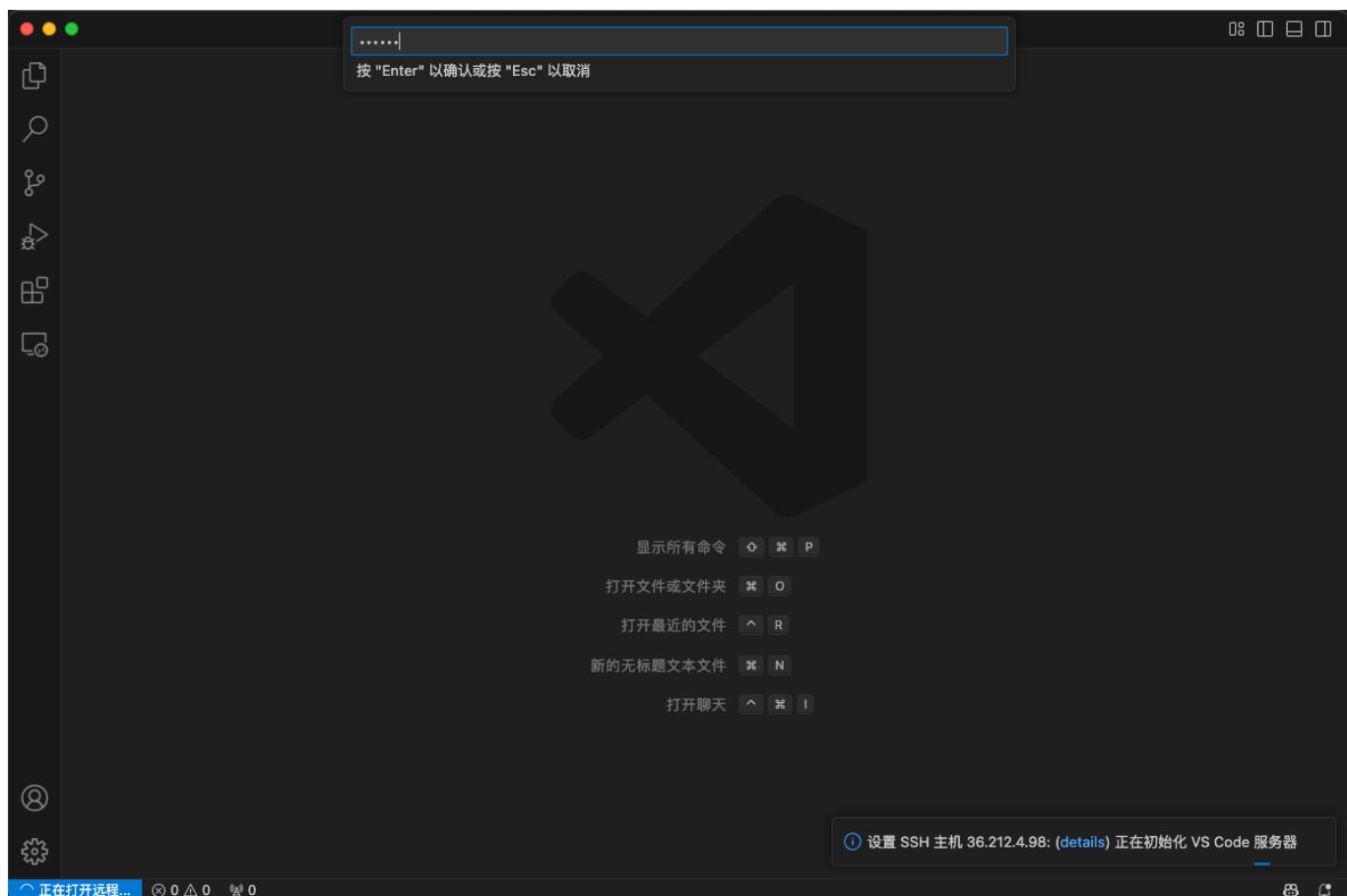
User tangou

```
config x  
Users > daylight > .ssh > config  
6 Host 36.212.4.98  
7 HostName 36.212.4.98  
8 User tangou  
9
```

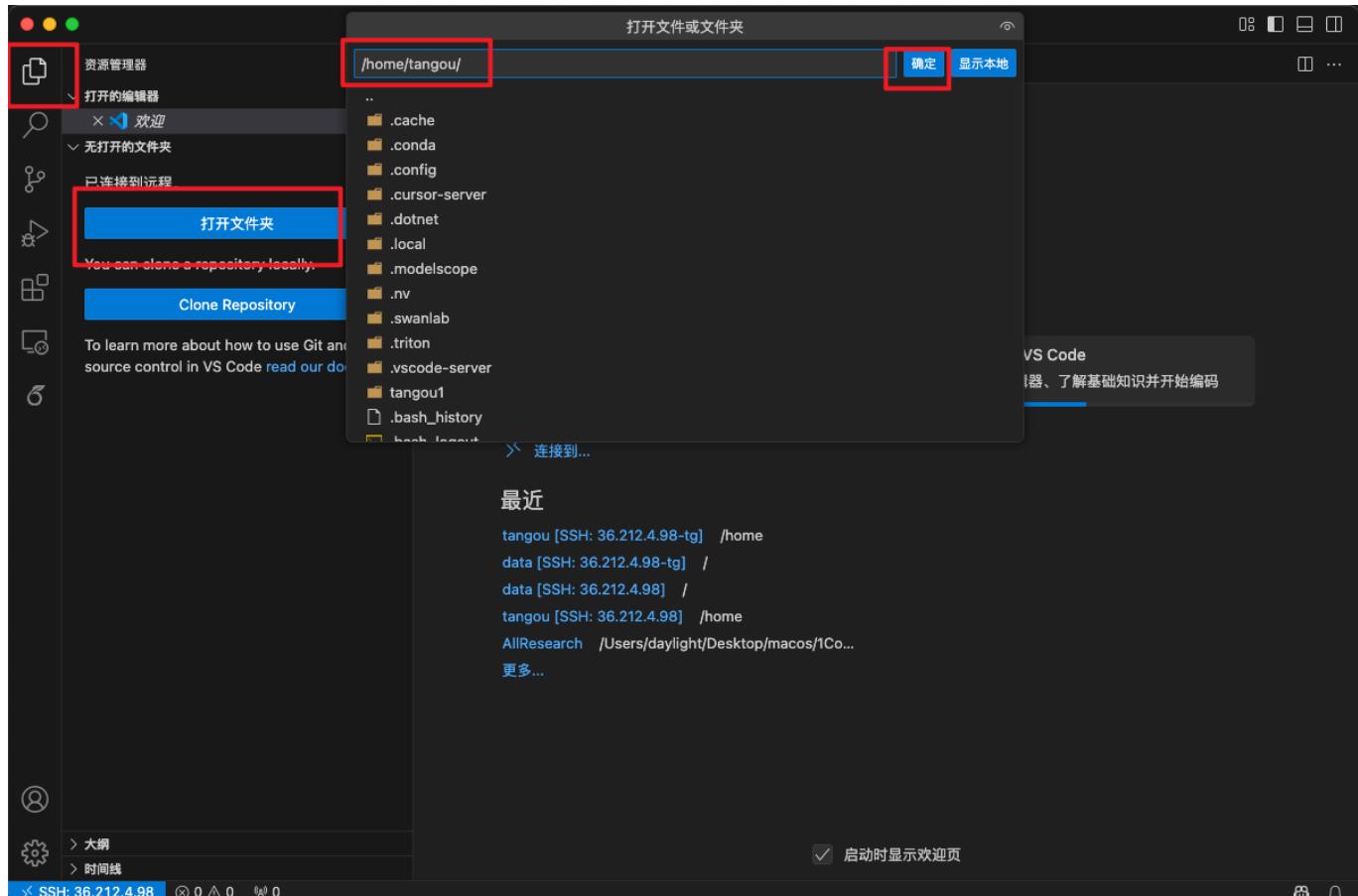
2. 点刷新，再打开文件



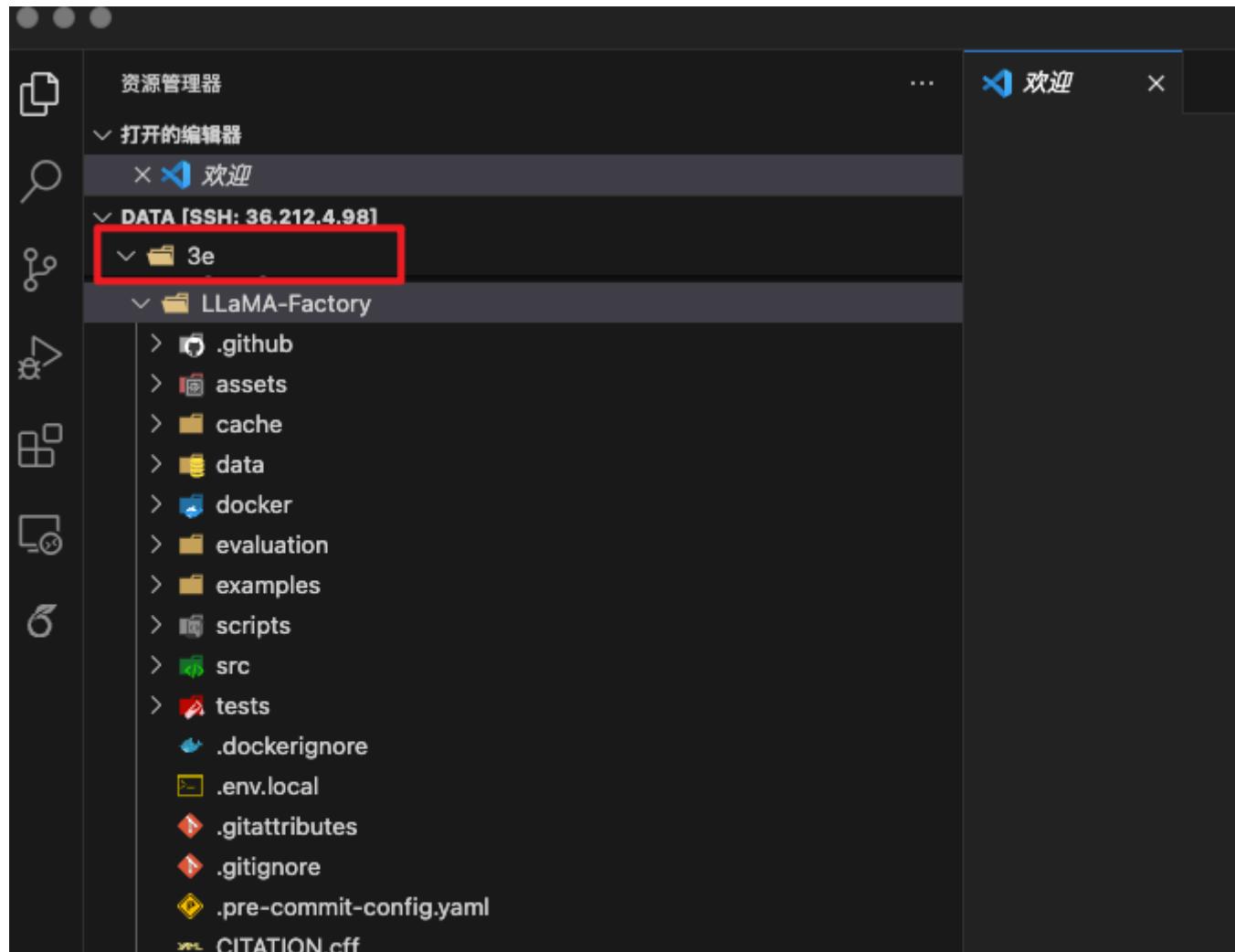
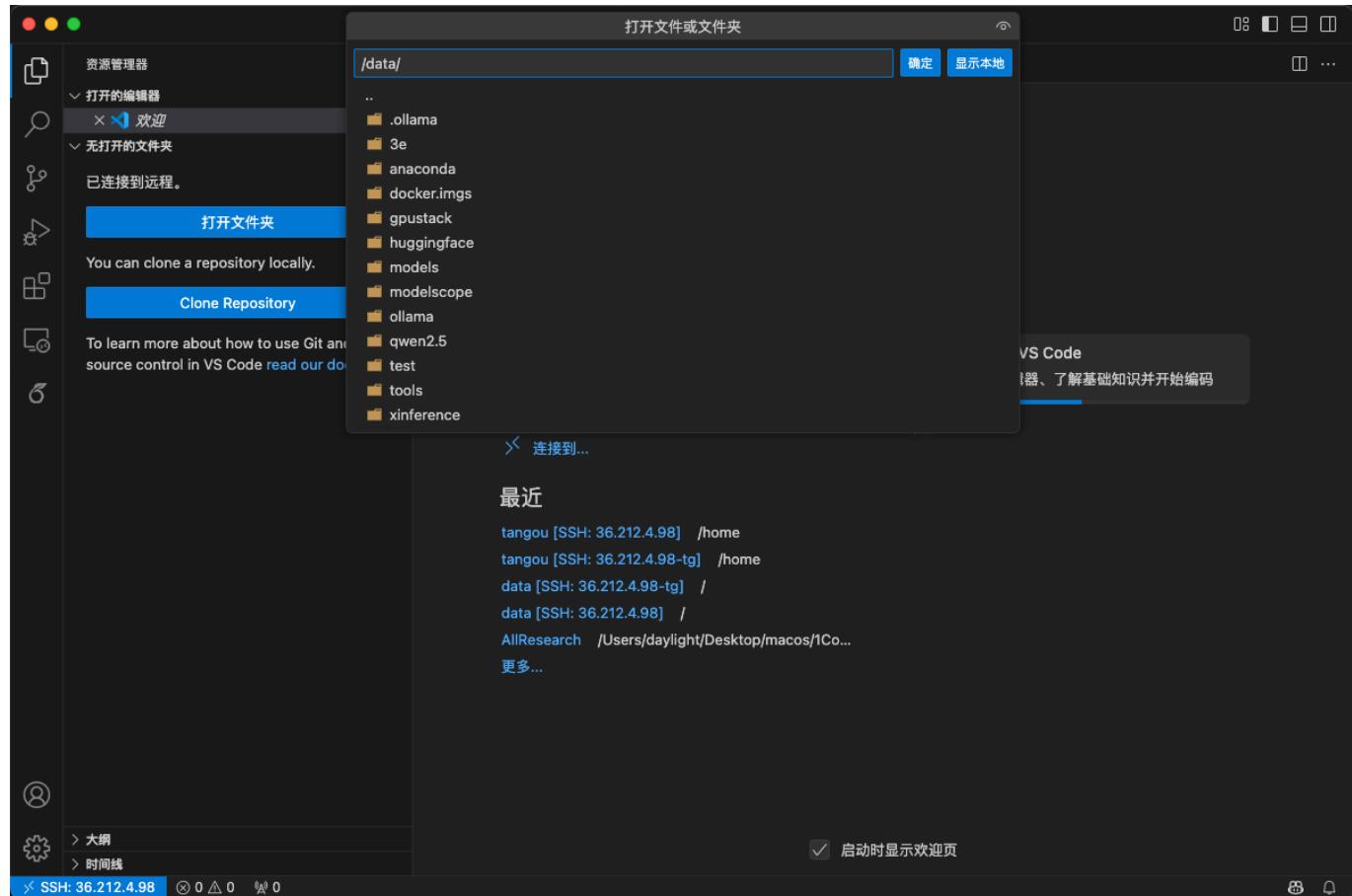
3. 输入密码123456，回车

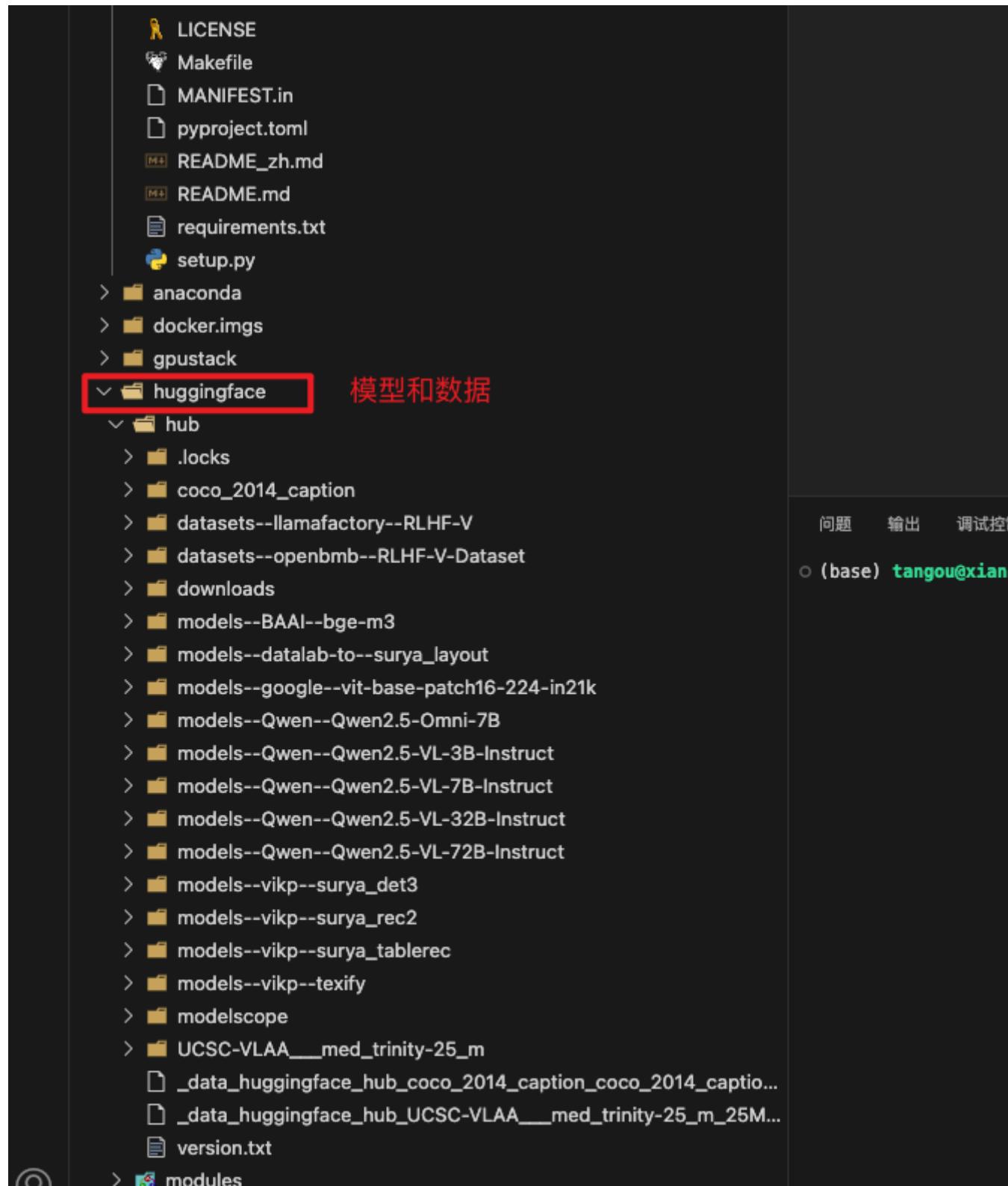


4. 进入



5. 回到第2步，再去开一个目录。这是共享模型数据的目录

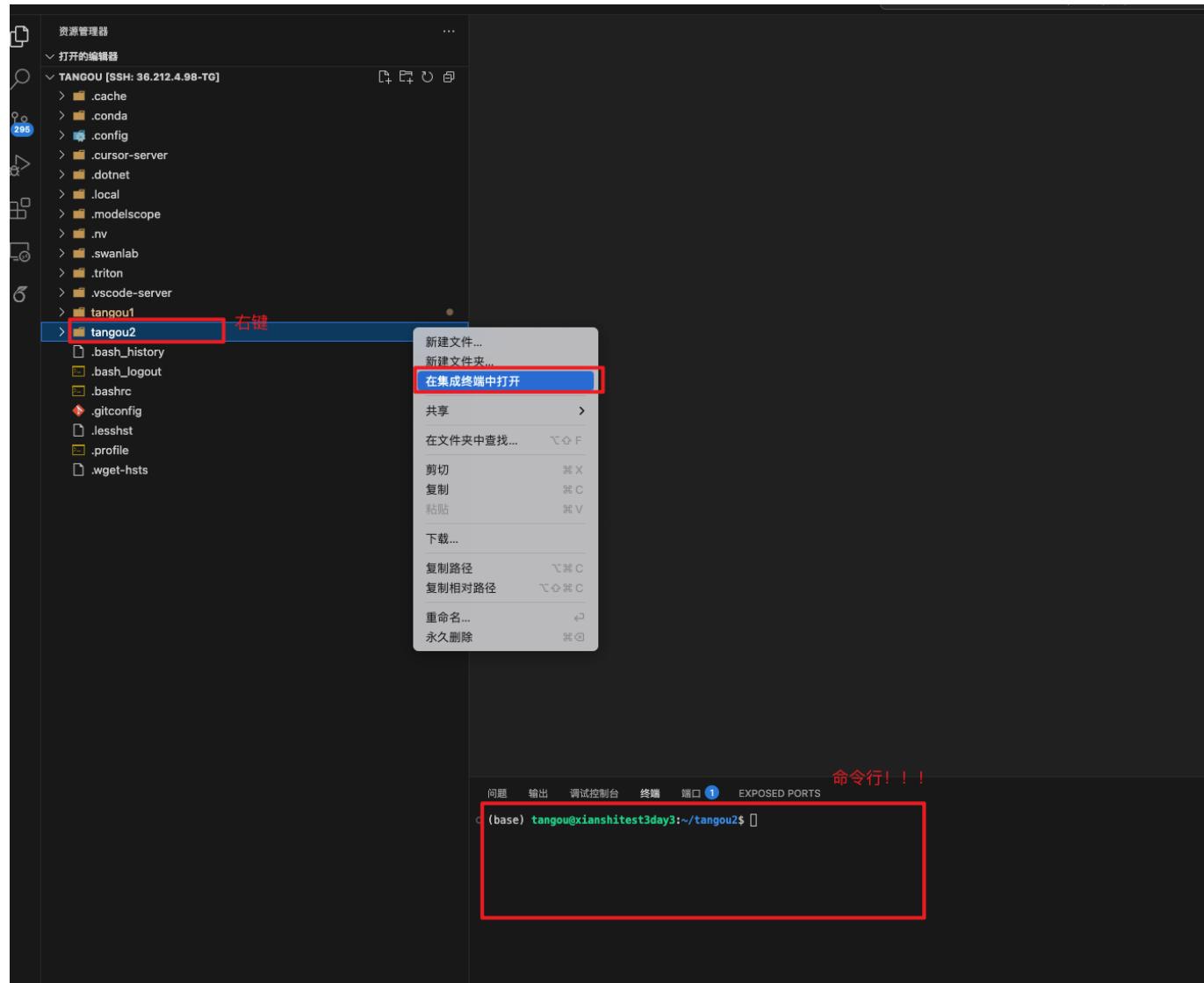




微调预备

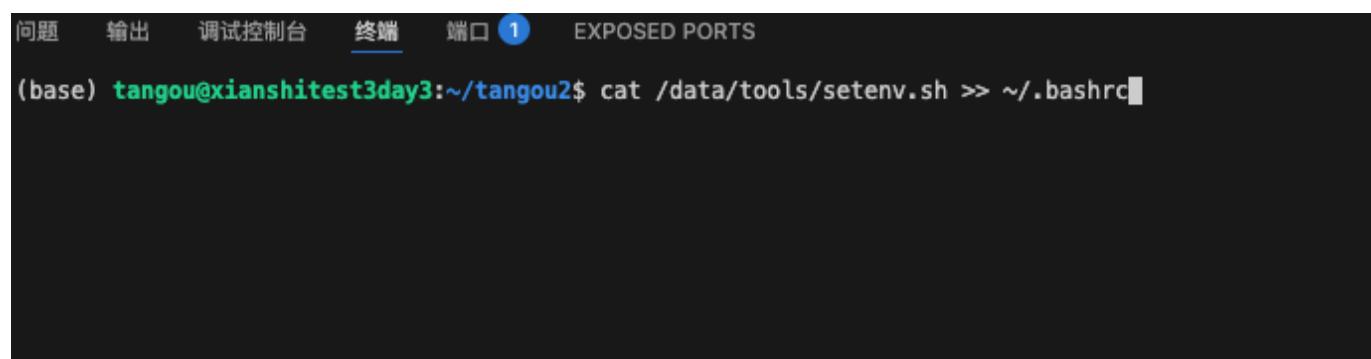
- 环境变量: conda、ollama

- 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
```



问题 输出 调试控制台 终端 端口 1 EXPOSED PORTS

- (base) tangou@xianshitest3day3:~/tangou2\$ cat /data/tools/setenv.sh >> ~/.bashrc^C
- (base) tangou@xianshitest3day3:~/tangou2\$ 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
```

```
(base) tangou@xianhitesh3day3:~/tangou$ conda info -envs 查看conda有哪些环境
# conda environments:
#
base                  * /data/anaconda
llama-factory          /data/anaconda/envs/llama-factory
llm                   /data/anaconda/envs/llm
pptagent               /data/anaconda/envs/pptagent
qwen2.5-omni            /data/anaconda/envs/qwen2.5-omni
qwen2.5-vl              /data/anaconda/envs/qwen2.5-vl
vtlm                  /data/anaconda/envs/vtlm

(base) tangou@xianhitesh3day3:~/tangou$ ollama list ollama查看模型
NAME           ID          SIZE    MODIFIED
qwen/qwen_5-vl-72b-instructlatest 8e7c8bcd35 145 GB  38 days ago
bytedance/textlatest 7307f64624807 121 GB  34 hours ago
bsahane/Qwen2.5-VL-7B-Instruct:04_K_M_benxh dcd488acac 4.7 GB  2 days ago
qwen2.5:32b 9f13ba1299af 19 GB   4 days ago
nonic-embed-text:latest 0a10f422947 274 MB  7 days ago
llmfusion/llmfusion:latest 08e5314d2c23 472 GB  7 days ago
qwen2.5-72b 424bb50cc137 497 GB  8 days ago
Huzder/deepseek-1-671b-2.5libit:latest b788bd59818 226 GB  12 days ago
deepseek-1:671b 739elbz29ad7 404 GB  12 days ago

(base) tangou@xianhitesh3day3:~/tangou$ ollama run bsahane/Qwen2.5-VL-7B-Instruct:04_K_M_benxh ollama运行其中一个模型
Hello! I'm here to help answer any questions you might have. Is there something specific you would like to know or discuss? I can provide information on a wide range of topics, from general knowledge to more detailed inquiries about various subjects.
Please feel free to ask anything that comes to mind!
```

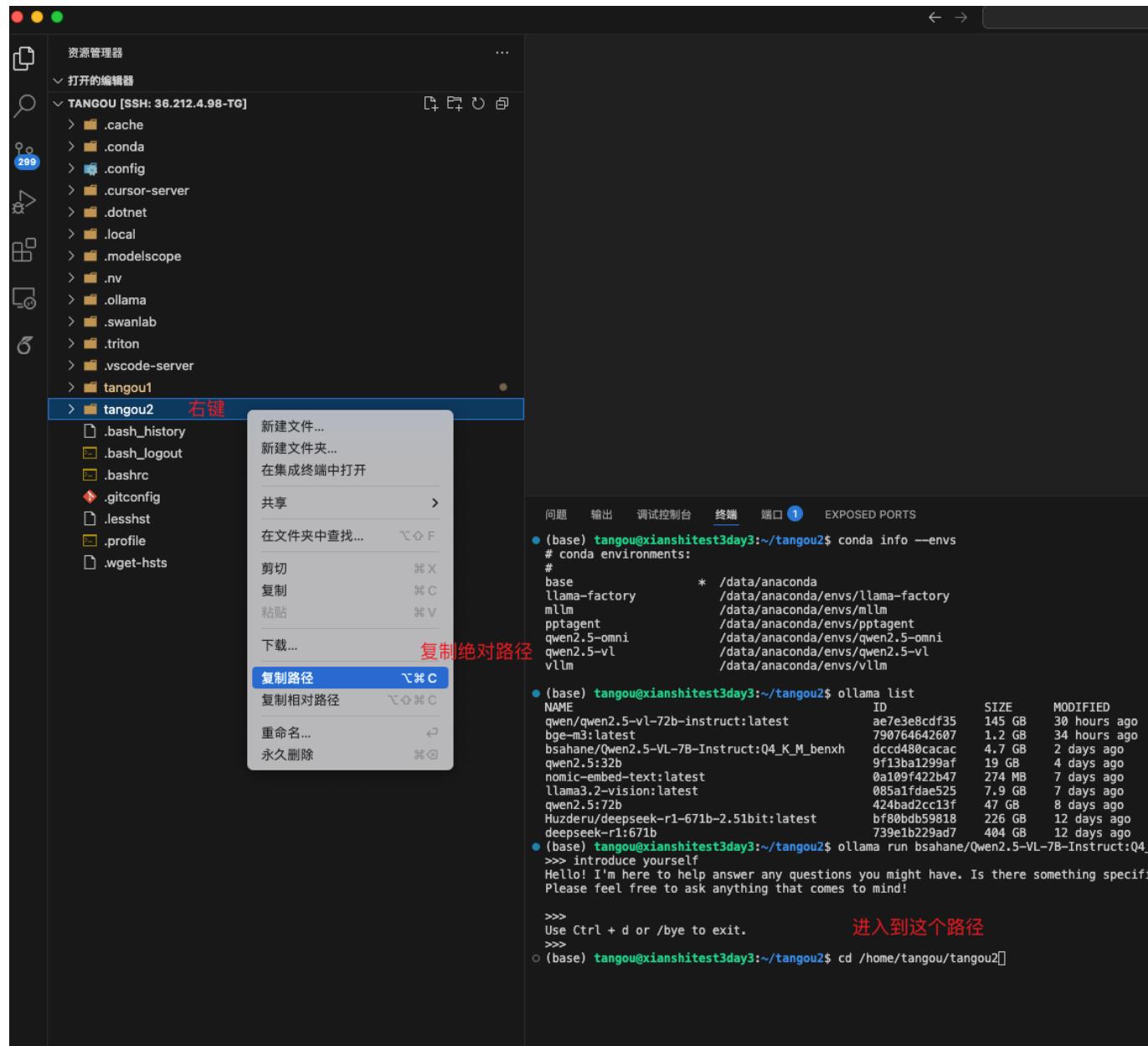
- vpn: 先不用管, 知道这个就行

```
# 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

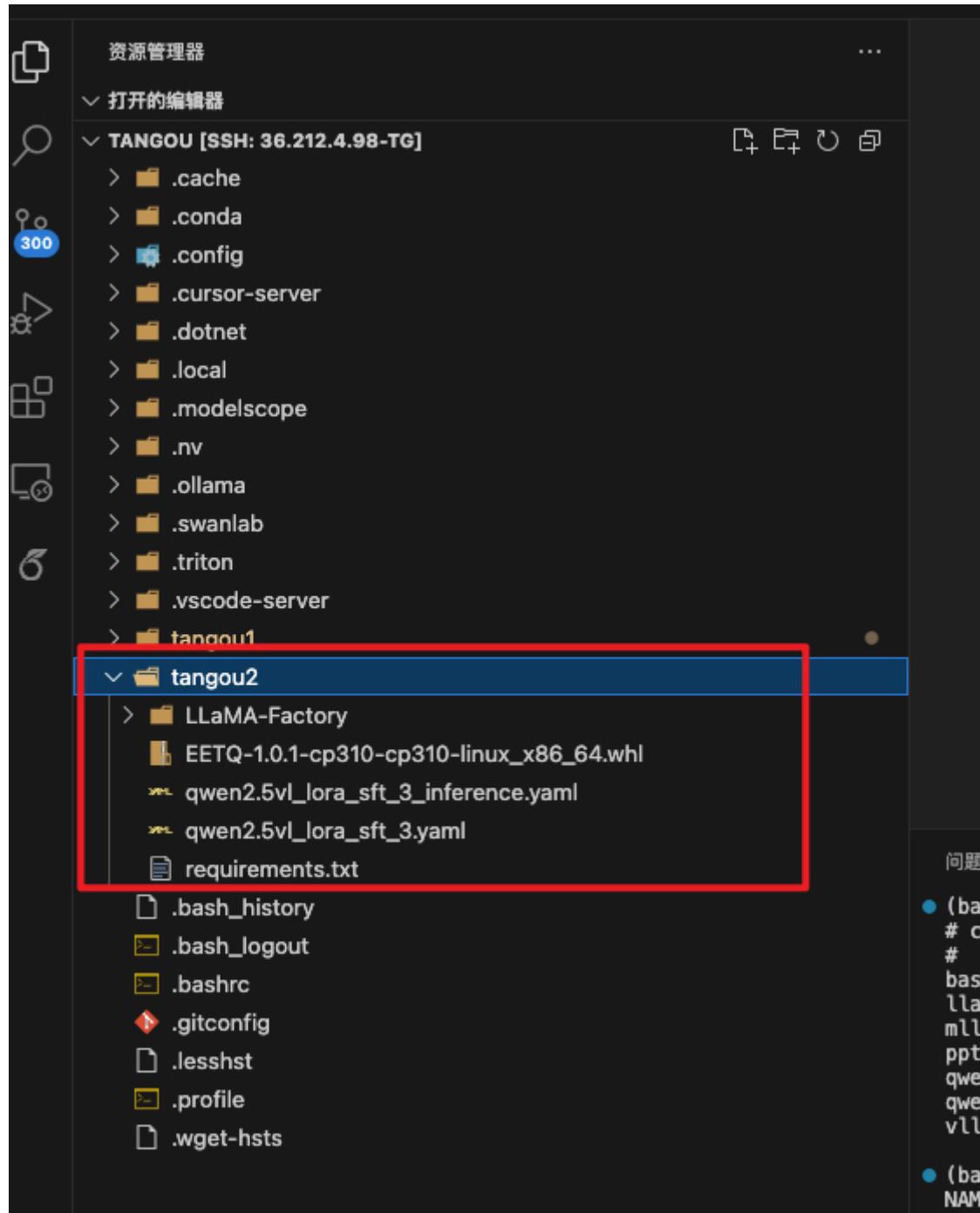
```
cd /home/tangou/tangou2 #你自己的路径
```



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
```

```

问题   输出   调试控制台   端口 ① EXPOSED PORTS
开vpn
(base) tangou@xianshitest3day3:~/tangou2$ source /data/tools/setproxy.sh
(base) tangou@xianshitest3day3:~/tangou2$ conda create -n tg10 python=3.10.16 创建自己的环境
Channels:
  - defaults
Platform: linux-64
Collecting package metadata (repodata.json): done
Solving environment: done
## Package Plan ##
environment location: /data/anaconda/envs/tg10
added / updated specs:
  - python=3.10.16

The following NEW packages will be INSTALLED:

libgcc_mutex      pkgs/main/linux-64::libgcc_mutex-0.1-main
openmp_mutex      pkgs/main/linux-64::openmp_mutex-5.1-l1_gnu
bz2                pkgs/main/linux-64::bz2-1.0.8-h5ee1bb_6
ca-certificates   pkgs/main/linux-64::ca-certificates-2025.2.25-h06a4308_0
ld_impl_linux-64  pkgs/main/linux-64::ld_impl_linux-64-2.40-h12ee557_0
libffi             pkgs/main/linux-64::libffi-3.4.4-hba678d5_1
libgcc-c-ng       pkgs/main/linux-64::libgcc_c-ng-3.4.4-h06a4308_0
libgcc            pkgs/main/linux-64::libgcc-3.4.4-h06a4308_0
libstdcxx-ng      pkgs/main/linux-64::libstdcxx-ng-11.2.0-h1234567_1
libuuid            pkgs/main/linux-64::libuuid-1.41.5-h5ee1bb_0
ncurses           pkgs/main/linux-64::ncurses-6.4-h6a78d5_0
openssl           pkgs/main/linux-64::openssl-3.0.16-h5ee1bb_0
pip               pkgs/main/linux-64::pip-25.0-py310h06a4308_0
python             pkgs/main/linux-64::python-3.10.16-h70216_1
readline           pkgs/main/linux-64::readline-8.2-h5ee1bb_0
sqlite             pkgs/main/linux-64::sqlite-3.45.3-h5ee1bb_0
tk                pkgs/main/linux-64::tk-8.6.14-h39e8969_0
tzdata            pkgs/main/noarch::tzdata-2025a-h04d1e81_0
wheel             pkgs/main/linux-64::wheel-0.45.1-py310h06a4308_0
xz                pkgs/main/linux-64::xz-5.6.4-h5ee1bb_1
zlib              pkgs/main/linux-64::zlib-1.2.13-h5ee1bb_1

Proceed ([y]/n)? y  输入y, 回车

Downloading and Extracting Packages:
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
#
# To activate this environment, use
# $ conda activate tg10
# To deactivate an active environment, use
# $ conda deactivate
(base) tangou@xianshitest3day3:~/tangou2$ conda activate tg10 切换环境
[tg10] tangou@xianshitest3day3:~/tangou2$ 这里发生变化, 成功

```

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

```

```

(tg10) tangou@xianshitest3day3:~/tangou2$ pip install -r requirements.txt
Collecting accelerate==1.2.1 (from -r requirements.txt (line 1))
  Using cached accelerate-1.2.1-py3-none-any.whl.metadata (19 kB)
Collecting adam_mini==1.1.1 (from -r requirements.txt (line 2))
  Using cached adam_mini-1.1.1-py3-none-any.whl.metadata (2.9 kB)
Collecting addict==2.4.0 (from -r requirements.txt (line 3))
  Using cached addict-2.4.0-py3-none-any.whl.metadata (1.0 kB)
Collecting aiofiles==23.2.1 (from -r requirements.txt (line 4))
  Using cached aiofiles-23.2.1-py3-none-any.whl.metadata (9.7 kB)
Collecting aiohappyeyeballs==2.6.1 (from -r requirements.txt (line 5))
  Using cached aiohappyeyeballs-2.6.1-py3-none-any.whl.metadata (5.9 kB)
Collecting aiohttp==3.11.13 (from -r requirements.txt (line 6))
  Using cached aiohttp-3.11.13-cp310-cp310-manylinux_2_17_x86_64.manylinux2014
Collecting aiosignal==1.3.2 (from -r requirements.txt (line 7))
  Using cached aiosignal-1.3.2-py2.py3-none-any.whl.metadata (3.8 kB)
Collecting airportsdata==20250224 (from -r requirements.txt (line 8))
  Using cached airportsdata-20250224-py3-none-any.whl.metadata (9.0 kB)
Collecting annotated-types==0.7.0 (from -r requirements.txt (line 9))
  Using cached annotated_types-0.7.0-py3-none-any.whl.metadata (15 kB)
Collecting anyio==4.8.0 (from -r requirements.txt (line 10))
  Using cached anyio-4.8.0-py3-none-any.whl.metadata (4.6 kB)
Collecting apollo-torch==1.0.3 (from -r requirements.txt (line 11))
  Using cached apollo_torch-1.0.3-py3-none-any.whl.metadata (15 kB)
Collecting aqlm==1.1.6 (from -r requirements.txt (line 12))
  Using cached aqlm-1.1.6-py3-none-any.whl.metadata (1.7 kB)

```

微调，这里用llamafactory提供的数据

1. 数据解读

资源管理器

打开的编辑器

TANGOU [SSH: 36.212.4.98-TG]

mlm_demo.json

LLaMA-Factory > data > mlm_demo.json

```
{ "messages": [ { "content": "<image>Who are they?", "role": "user" }, { "content": "They're Kane and Gretzka from Bayern Munich.", "role": "assistant" }, { "content": "What are they doing?<image>", "role": "user" }, { "content": "They are celebrating on the soccer field.", "role": "assistant" } ], "images": [ "mlm_demo_data/1.jpg", "mlm_demo_data/1.jpg" ] }, { "messages": [ { "content": "<image>Who is he?", "role": "user" }, { "content": "He's Thomas Muller from Bayern Munich.", "role": "assistant" }, { "content": "Why is he on the ground?", "role": "user" }, { "content": "Because he's sliding on his knees to celebrate.", "role": "assistant" } ], "images": [ "mlm_demo_data/2.jpg" ] }
```

问题 输出 调试控制台 终端 端口 EXPOSED PORTS

(tg10) tangou@xianshi-test3day3:~/tangou2/LLaMA-Factory\$ llmfactory-cli webui
[2025-03-31 21:11:53,407] [INFO] [real_accelerator.py:22: get_accelerator] Setting ds_accelerator to cuda (auto)
* Running on local URL: http://0.0.0.0:7860

To create a public link, set `share=True` in `launch()`.

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

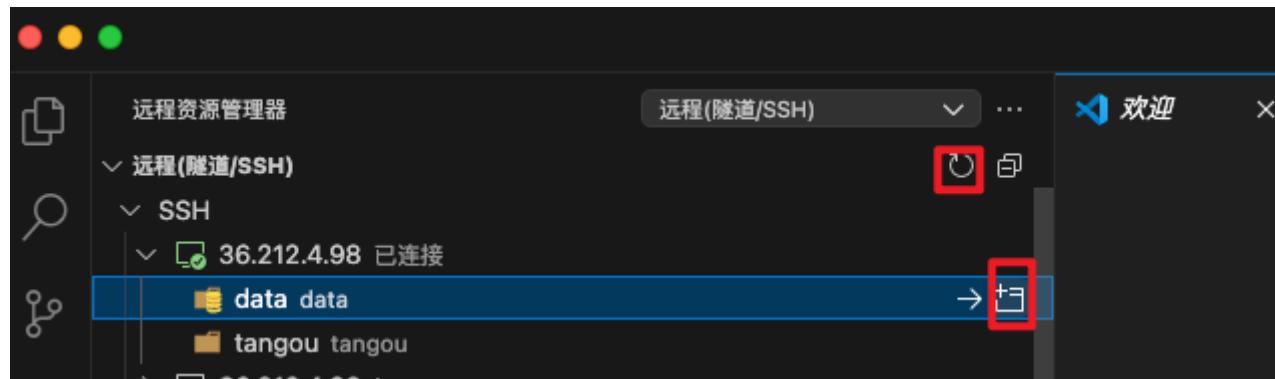
The screenshot shows a terminal window with two tabs open. The left tab contains the JSON file `mlm_demo.json`, which defines a dataset structure. The right tab contains the configuration file `dataset_info.json`. A red box highlights the `dataset_info.json` file.

```
mlm_demo.json
{
    "dataset": {
        "file_name": "mlm_demo.json",
        "columns": [
            "text", "messages"
        ],
        "images": [
            "images"
        ],
        "audios": [
            "audios"
        ],
        "videos": [
            "videos"
        ]
    },
    "tags": [
        {
            "role_tag": "role",
            "content_tag": "content",
            "user_tag": "user",
            "assistant_tag": "assistant"
        }
    ],
    "audio_demo": [
        {
            "file_name": "mlm_audio_demo.json",
            "formatting": "sharept",
            "columns": [
                "messages"
            ],
            "audios": [
                "audios"
            ],
            "tags": [
                {
                    "role_tag": "role",
                    "content_tag": "content",
                    "user_tag": "user",
                    "assistant_tag": "assistant"
                }
            ]
        }
    ],
    "video_demo": [
        {
            "file_name": "mlm_video_demo.json",
            "formatting": "sharept",
            "columns": [
                "messages"
            ],
            "videos": [
                "videos"
            ],
            "tags": [
                {
                    "role_tag": "role",
                    "content_tag": "content",
                    "user_tag": "user",
                    "assistant_tag": "assistant"
                }
            ]
        }
    ],
    "alpaca_en": {
        "hf_hub_url": "l1amafactory/alpaca_en",
        "mlm_hub_url": "l1amafactory/alpaca_en"
    }
}
```

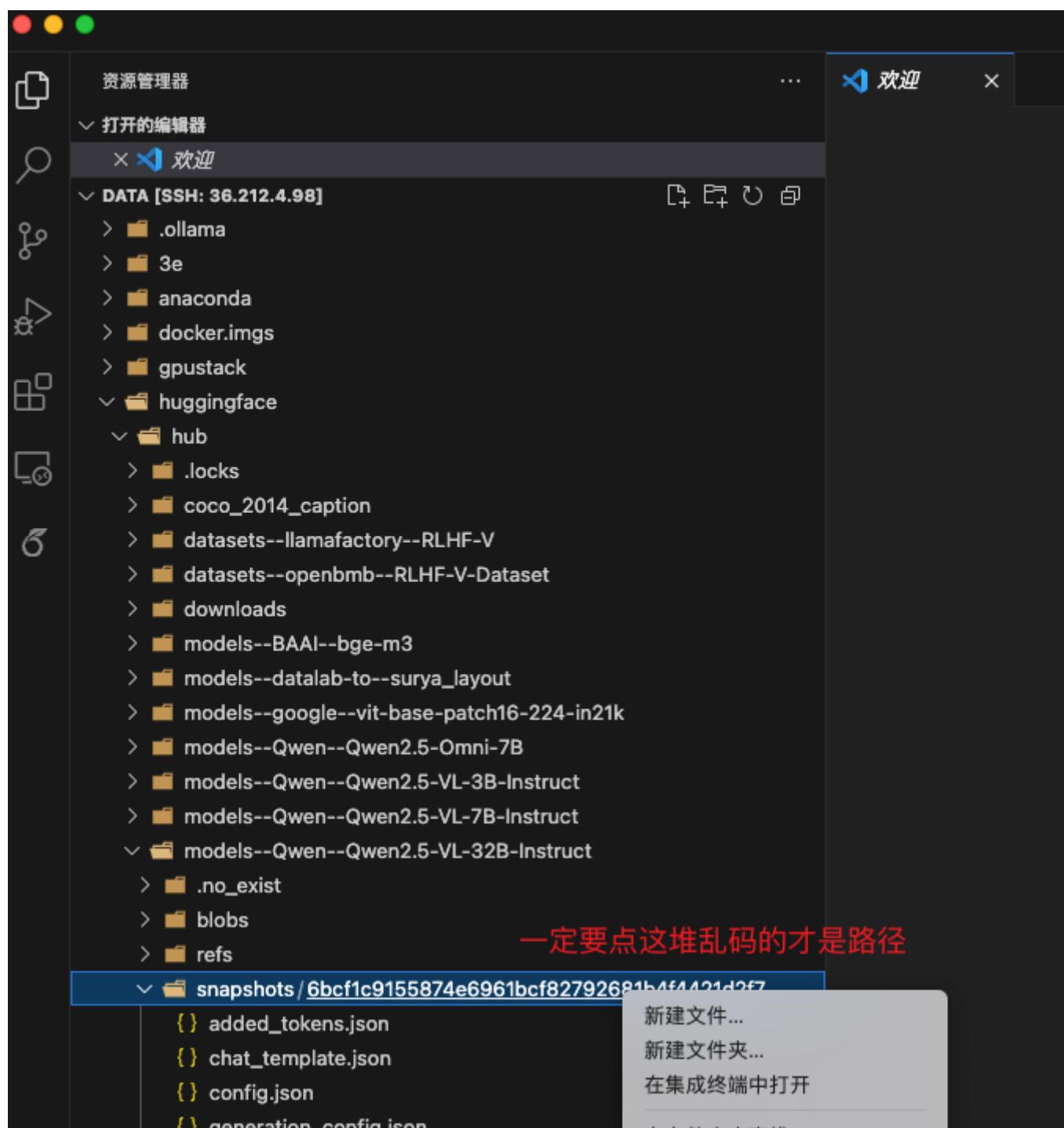
```
dataset_info.json
{
    "file_name": "mlm_demo.json",
    "columns": [
        "text", "messages"
    ],
    "images": [
        "images"
    ],
    "audios": [
        "audios"
    ],
    "videos": [
        "videos"
    ]
},
{
    "tags": [
        {
            "role_tag": "role",
            "content_tag": "content",
            "user_tag": "user",
            "assistant_tag": "assistant"
        }
    ],
    "audio_demo": [
        {
            "file_name": "mlm_audio_demo.json",
            "formatting": "sharept",
            "columns": [
                "messages"
            ],
            "audios": [
                "audios"
            ],
            "tags": [
                {
                    "role_tag": "role",
                    "content_tag": "content",
                    "user_tag": "user",
                    "assistant_tag": "assistant"
                }
            ]
        }
    ],
    "video_demo": [
        {
            "file_name": "mlm_video_demo.json",
            "formatting": "sharept",
            "columns": [
                "messages"
            ],
            "videos": [
                "videos"
            ],
            "tags": [
                {
                    "role_tag": "role",
                    "content_tag": "content",
                    "user_tag": "user",
                    "assistant_tag": "assistant"
                }
            ]
        }
    ],
    "alpaca_en": {
        "hf_hub_url": "l1amafactory/alpaca_en",
        "mlm_hub_url": "l1amafactory/alpaca_en"
    }
}
```

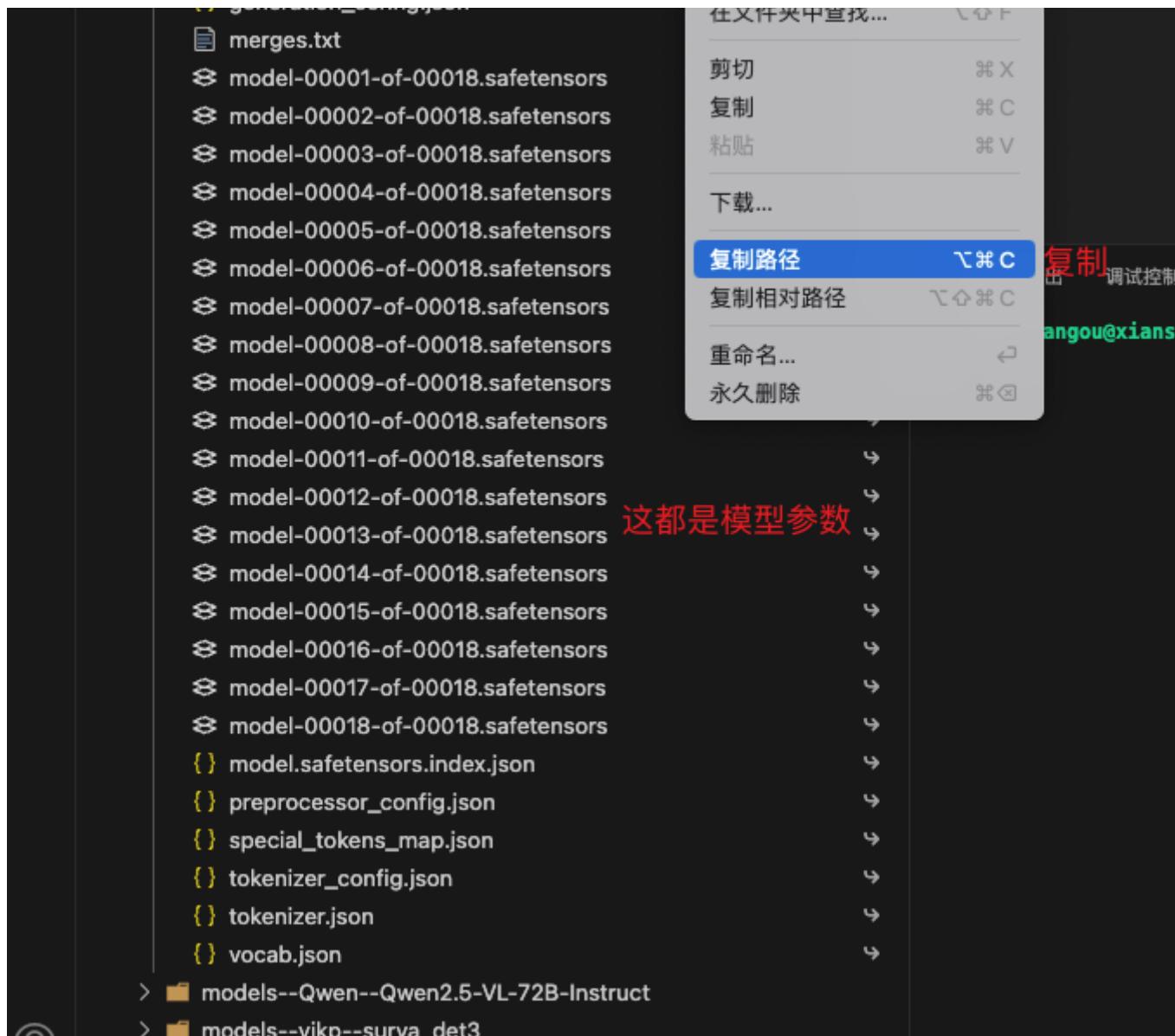
3. 配置模型路径

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

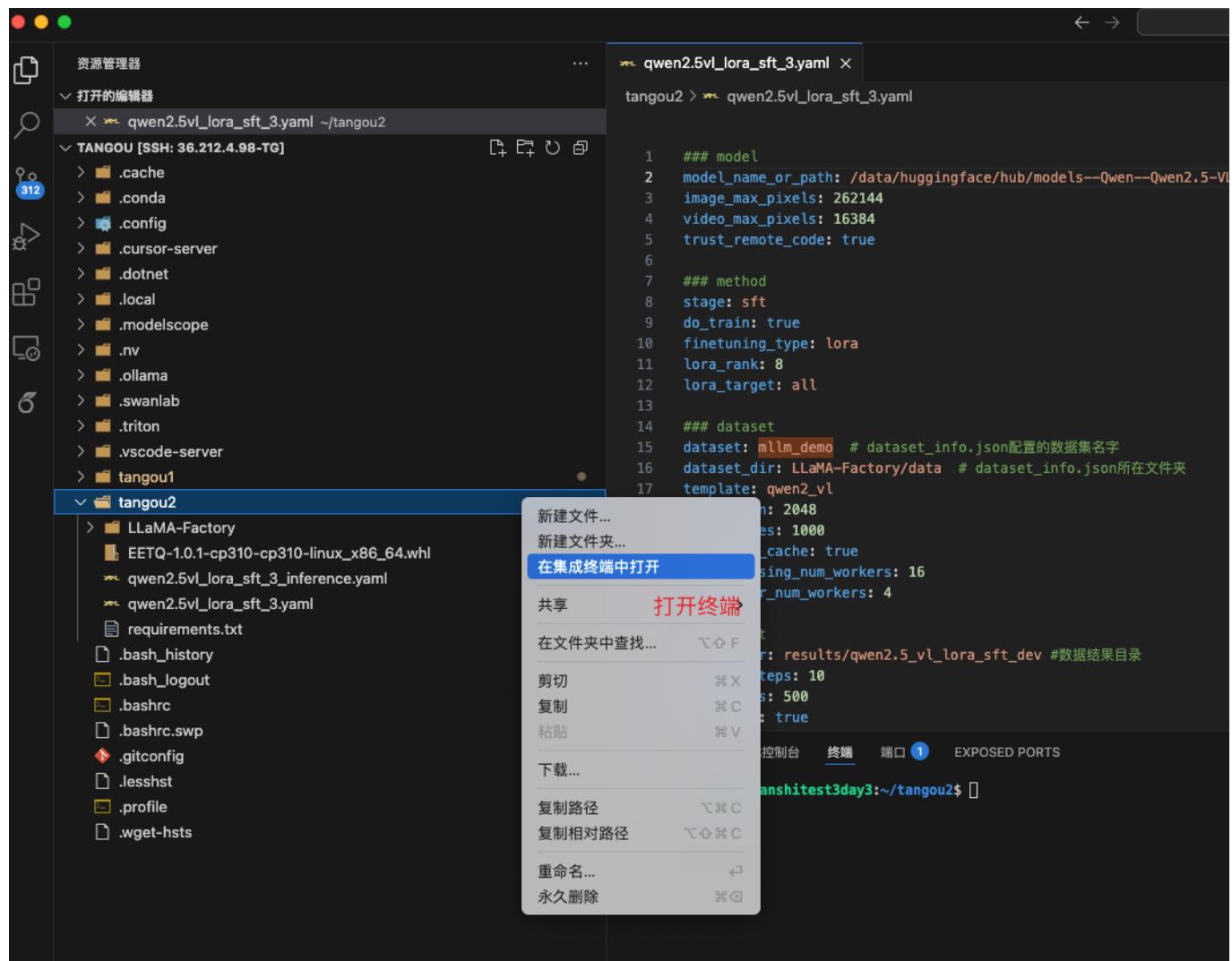
```

#> tangou2 > qwen2.5vl_lora_sft_3.yaml
1 ## model
2 model_name_or_path: /data/huggingface/hub/models--Qwen--Qwen2.5-VL-32B-Instruct/snapshots/6bcf1c9155874e6961bc182792681b4f442102f #配置model路径, 回到最开始3e那个自带去copy路径
3 max_seq_length: 2048
4 video_max_pixels: 16384
5 trust_remote_code: true
6
7 ### method
8 stage: sft
9 do_train: true
10 finetuning_type: lora
11 lora_rank: 8
12 lora_target: all
13
14 ## dataset
15 dataset: llm_demo # dataset_info.json配置的数据集名字
16 dataset_dir: LLaMA-Factory/data # dataset_info.json所在文件夹
17 compressed_llm_vocab
18 cutoff_len: 2048
19 max_samples: 1000
20 overwrite_cache: true
21 preprocessing_num_workers: 16
22 dataloader_num_workers: 4
23
24 ##> output
25 output_dir: results/qwen2.5_vl_lora_sft_dev #数据结果目录
26 logging_steps: 10
27 save_steps: 500
28 plot_loss: true

```

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 llmfactory-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
```

```
tangou2 ~ % qwen2.5vl_lora_sft_3.yaml ~/tangou2
tangou2 ~ % qwen2.5vl_lora_sft_3.yaml

1 ##### model
2 model_name_or_path: /data/huggingface/hub/models—Qwen—Qwen-2.5-VL-32B-Instruct/snapshots/[redacted] #配置model路径, 回到最开始3e那个目录去copy路径
3 image_max_pixels: 262144
4 video_max_pixels: 16384
5 trust_remote_code: true
6
7 ##### method
8 stage: sft
9 do_train: true
10 finetuning_type: lora
11 lora_rank: 8
12 lora_target: all
13
14 ##### dataset
15 dataset: plm_dmp # dataset_info.json配置的数据集名字
16 dataset_dir: LLaMA-Factory/data # dataset_info.json所在文件夹
17 template: qwen2.vl
18 cutoff_len: 2048
19 max_samples: 1000
20 overwrite_cache: true
21 preprocessing_num_workers: 16
22 dataloader_num_workers: 4
23
24 ##### output
25 output_dir: results/qwen2.5_vl_lora_sft_dev #数据结果目录
26 logging_steps: 10
27 save_steps: 500
28 plot_loss: true

问题 输出 测试控制台 终端 窗口 ① EXPOSED PORTS

(tg10) tangouxianshi test3d@3d3:~/tangou2$ source /data/tools/setproxy.sh
(tg10) tangouxianshi test3d@3d3:~/tangou2$ conda activate tg10
(tg10) tangouxianshi test3d@3d3:~/tangou2$ NCCL_P2P_LEVEL=4V1 HUGGINGFACE_HUB_CACHE="/data/huggingface/hub" FORCE_TORCHRUN=1 CUDA_VISIBLE_DEVICES=4,5,6,7 llamafactory-cli train qwen2.5vl_lora_sft_3.yaml
这里一定要切换成功
```

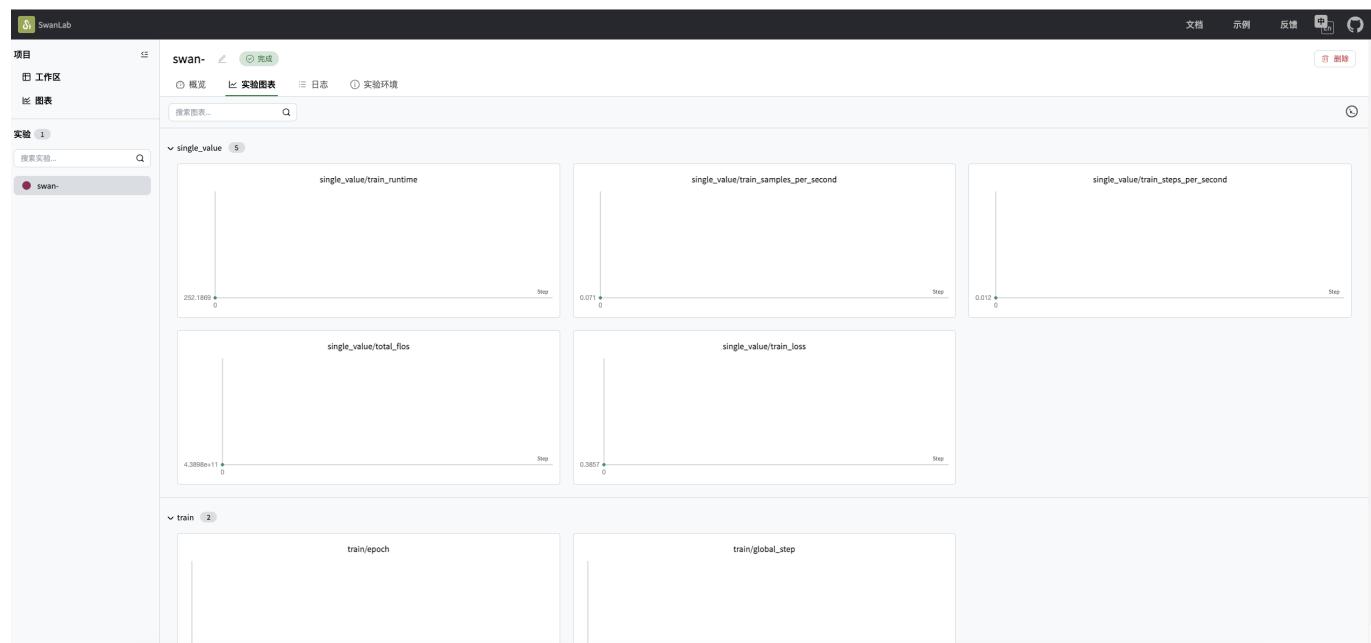
- > 📂 .conda
- > 📂 .config
- > 📂 .cursor-server
- > 📂 .dotnet
- > 📂 .local
- > 📂 .modelscope
- > 📂 .nv
- > 📂 .ollama
- > 📂 .swanlab
- > 📂 .triton
- > 📂 .vscode-server
- > 📂 tangou1
- ✓ 📂 tangou2
 - > 📂 data
 - > 📂 LLaMA-Factory
 - > 📂 results
 - > 📂 swanlog
 - ─ EETQ-1.0.1-cp310-cp310-linux_x86_64.whl
 - ─ qwen2.5vl_lora_sft_3_inference.yaml
 - ─ qwen2.5vl_lora_sft_3.yaml
 - ─ requirements.txt
- ─ .bash_history
- ─ .bash_logout
- ─ .bashrc

- (tg10) tangou@xianshitest3day3:~/tangou2\$ swanlab watch swanlog
swanboard: SwanLab Experiment Dashboard is up-to-date in time

→ Local: http://127.0.0.1:5092
→ press **ctrl + F** to pull

问题	输出	调试控制台	终端	端口 2	EXPOSED PORTS	正
				端口	转发地址	
<input type="radio"/> 5092					127.0.0.1:5092	
<input type="radio"/> 7860					localhost:7860	
添加端口						
映射下端口5092，然后去本地浏览器输入127.0.0.1:5092						

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 Usage	
ID	ID						
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	

Tue Apr 1 01:22:49 2025					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	38505MiB / 40960MiB	0%	Default	0	Disabled	
N/A	34C	P0	37W / 250W						
1	NVIDIA A100-PCIE-40GB	Off	00000000:3E:00.0 Off	38505MiB / 40960MiB	0%	Default	0	Disabled	
N/A	34C	P0	38W / 250W						
2	NVIDIA A100-PCIE-40GB	Off	00000000:40:00.0 Off	38505MiB / 40960MiB	0%	Default	0	Disabled	
N/A	33C	P0	38W / 250W						
3	NVIDIA A100-PCIE-40GB	Off	00000000:41:00.0 Off	38505MiB / 40960MiB	0%	Default	0	Disabled	
N/A	35C	P0	40W / 250W						
4	NVIDIA A100-PCIE-40GB	Off	00000000:B1:00.0 Off	38505MiB / 40960MiB	0%	Default	0	Disabled	
N/A	34C	P0	36W / 250W						
5	NVIDIA A100-PCIE-40GB	Off	00000000:B2:00.0 Off	38505MiB / 40960MiB	0%	Default	0	Disabled	
N/A	35C	P0	41W / 250W						
6	NVIDIA A100-PCIE-40GB	Off	00000000:B4:00.0 Off	38505MiB / 40960MiB	0%	Default	0	Disabled	
N/A	34C	P0	36W / 250W						
7	NVIDIA A100-PCIE-40GB	Off	00000000:B5:00.0 Off	38505MiB / 40960MiB	0%	Default	0	Disabled	
N/A	34C	P0	37W / 250W						

Processes:								
GPU	GI	CI	PID	Type	Process name	GPU Memory Usage		
ID		ID						
0	N/A	N/A	1974492	C	...anaconda/envs/tg10/bin/python	38488MiB		
1	N/A	N/A	1977989	C	...anaconda/envs/tg10/bin/python	38488MiB		
2	N/A	N/A	1977990	C	...anaconda/envs/tg10/bin/python	38488MiB		
3	N/A	N/A	1977991	C	...anaconda/envs/tg10/bin/python	38488MiB		
4	N/A	N/A	1977992	C	...anaconda/envs/tg10/bin/python	38488MiB		
5	N/A	N/A	1977993	C	...anaconda/envs/tg10/bin/python	38488MiB		
6	N/A	N/A	1977994	C	...anaconda/envs/tg10/bin/python	38488MiB		
7	N/A	N/A	1977995	C	...anaconda/envs/tg10/bin/python	38488MiB		

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
```

```
qwen2.5vl_lora_sft_3.yaml  -- qwen2.5vl_lora_sft_3_inference.yaml
tangou2 > -- qwen2.5vl_lora_sft_3_inference.yaml
1 model_name_or_path: data/huggingface/hub/models--Qwen--Qwen2.5-VL-32B-Instruct/snapshots/8bcf1c9155874e6961bcf82792681b4f4421d2f7
2 adapter_name_or_path: results/qwen2.5_vl_lora_sft_dev
3 template: qwen2
4 infer_backend:vllm # choices: [huggingface, vllm]
5 trust_remote_code: true
```

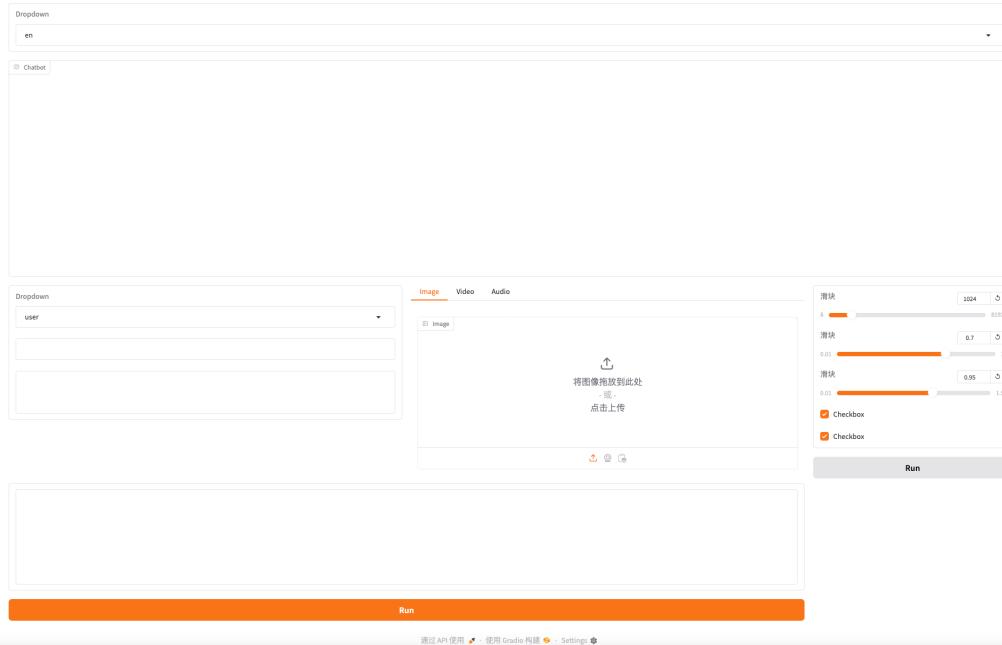
问题 输出 调试控制台 端口 EXPOSED PORTS

(tg@) tangouxianshitest3day1:~/tangou2\$ 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

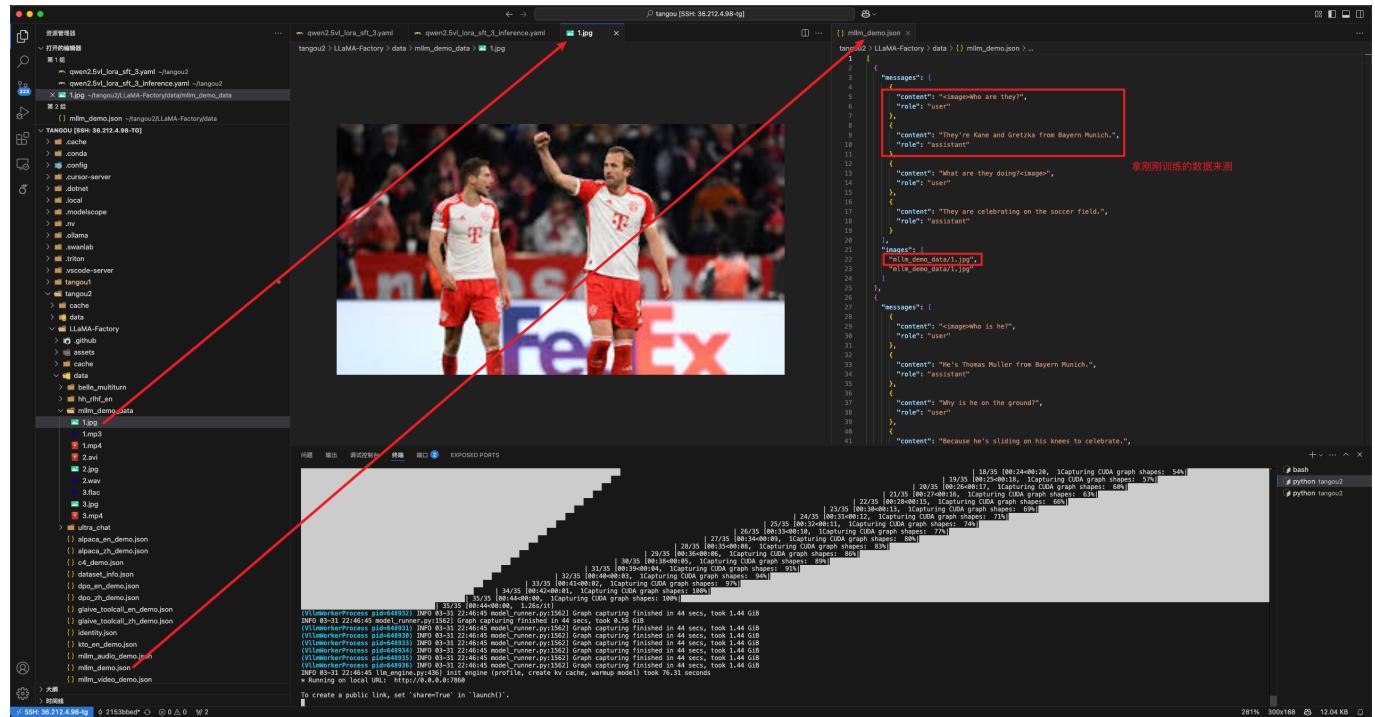
问题 输出 调试控制台 端口 ② EXPOSED PORTS

```
Capturing CUDA graph shapes: 60%
Capturing CUDA graph shapes: 63%
Capturing CUDA graph shapes: 66%
Capturing CUDA graph shapes: 69%
Capturing CUDA graph shapes: 71%
Capturing CUDA graph shapes: 74%
Capturing CUDA graph shapes: 77%
Capturing CUDA graph shapes: 80%
Capturing CUDA graph shapes: 83%
Capturing CUDA graph shapes: 86%
Capturing CUDA graph shapes: 89%
Capturing CUDA graph shapes: 91%
Capturing CUDA graph shapes: 94%
Capturing CUDA graph shapes: 97%
Capturing CUDA graph shapes: 100%
Capturing CUDA graph shapes: 100%[.26s]
[VLMWorkerProcess pid=648932] INFO 03-31 22:46:45 model_runner.py:1562] Graph capturing finished in 44 secs, took 1.44 GiB
INFO 03-31 22:46:45 model_runner.py:1562] Graph capturing finished in 44 secs, took 0.58 GiB
[VLMWorkerProcess pid=648931] INFO 03-31 22:46:45 model_runner.py:1562] Graph capturing finished in 44 secs, took 1.44 GiB
[VLMWorkerProcess pid=648930] INFO 03-31 22:46:45 model_runner.py:1562] Graph capturing finished in 44 secs, took 1.44 GiB
[VLMWorkerProcess pid=648933] INFO 03-31 22:46:45 model_runner.py:1562] Graph capturing finished in 44 secs, took 1.44 GiB
[VLMWorkerProcess pid=648934] INFO 03-31 22:46:45 model_runner.py:1562] Graph capturing finished in 44 secs, took 1.44 GiB
[VLMWorkerProcess pid=648935] INFO 03-31 22:46:45 model_runner.py:1562] Graph capturing finished in 44 secs, took 1.44 GiB
[VLMWorkerProcess pid=648936] INFO 03-31 22:46:45 model_runner.py:1562] Graph capturing finished in 44 secs, took 1.44 GiB
INFO 03-31 22:46:45 llm_engine_nv436l_init_engine (profile, create_kv_cache, warmup_model) took 76.31 seconds
* Running on local URL: http://0.0.0.0:7860 一般会自动映射端口, 没映射的话去映射下
To create a public link, set `share=True` in `launch()`.
```

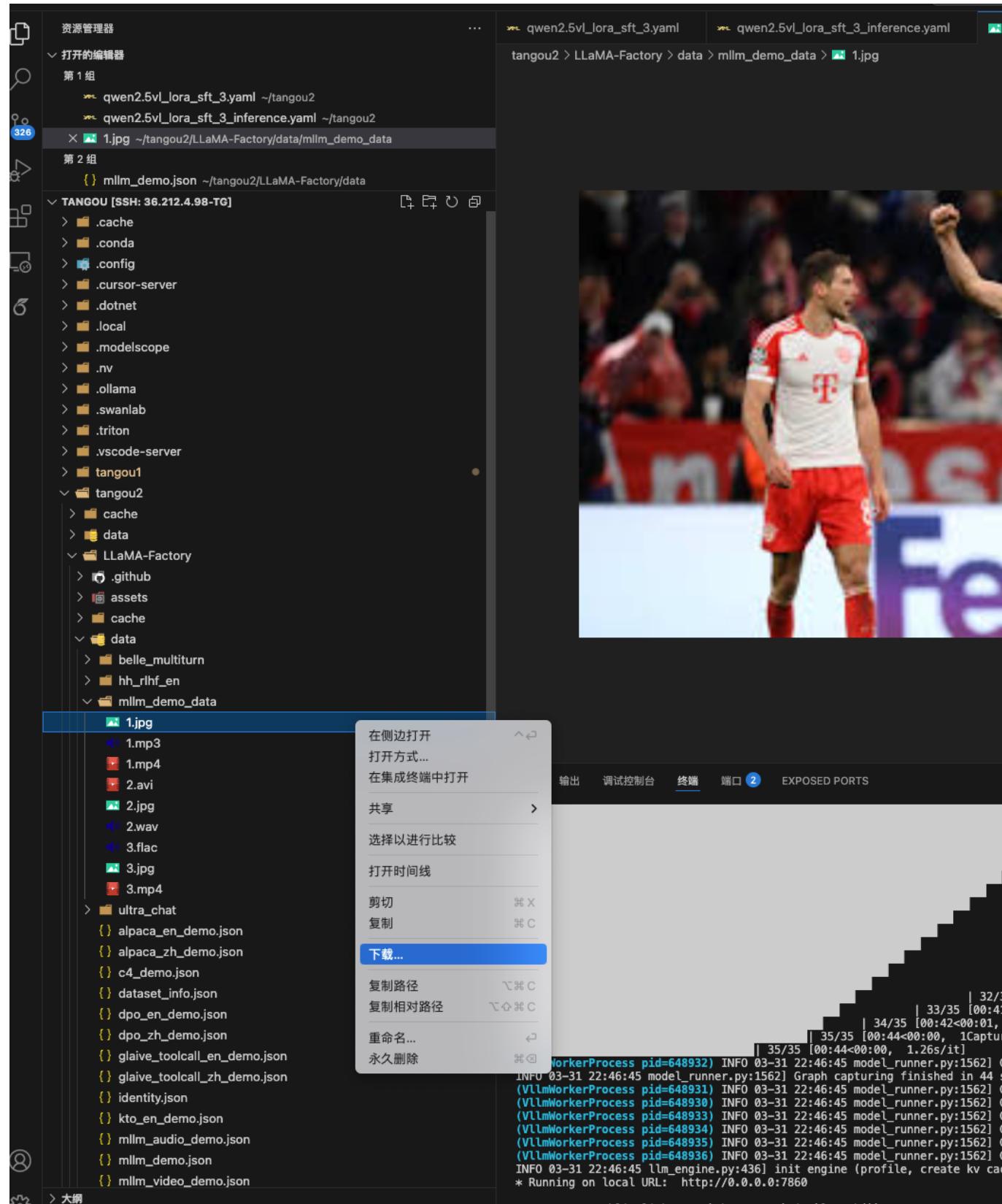
本地浏览器访问: <http://0.0.0.0:7860>



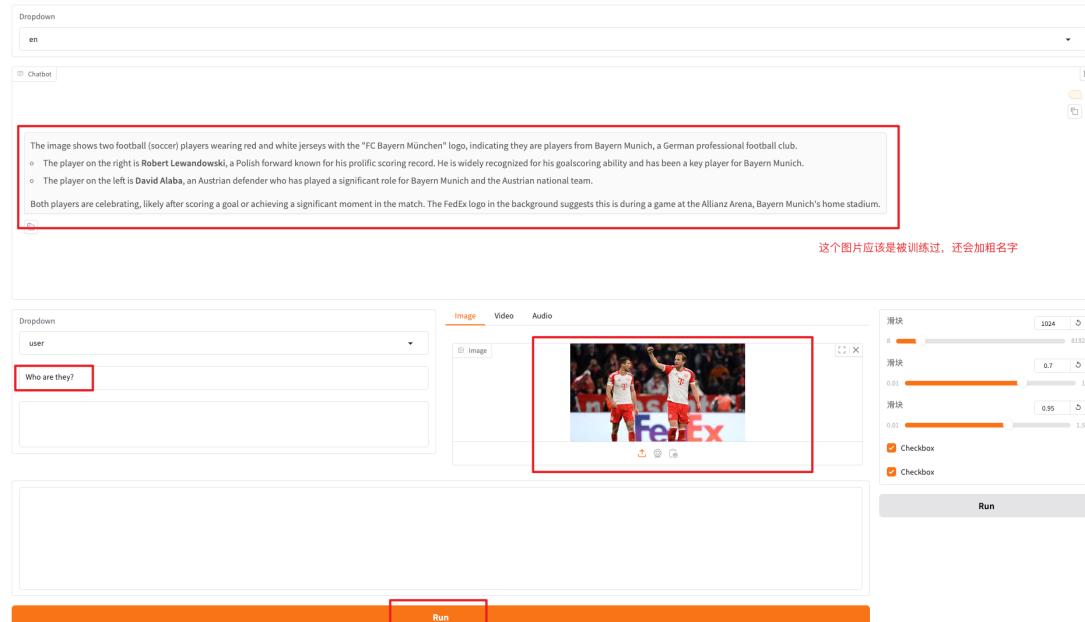
拿刚刚训练的数据来测



下载到本地

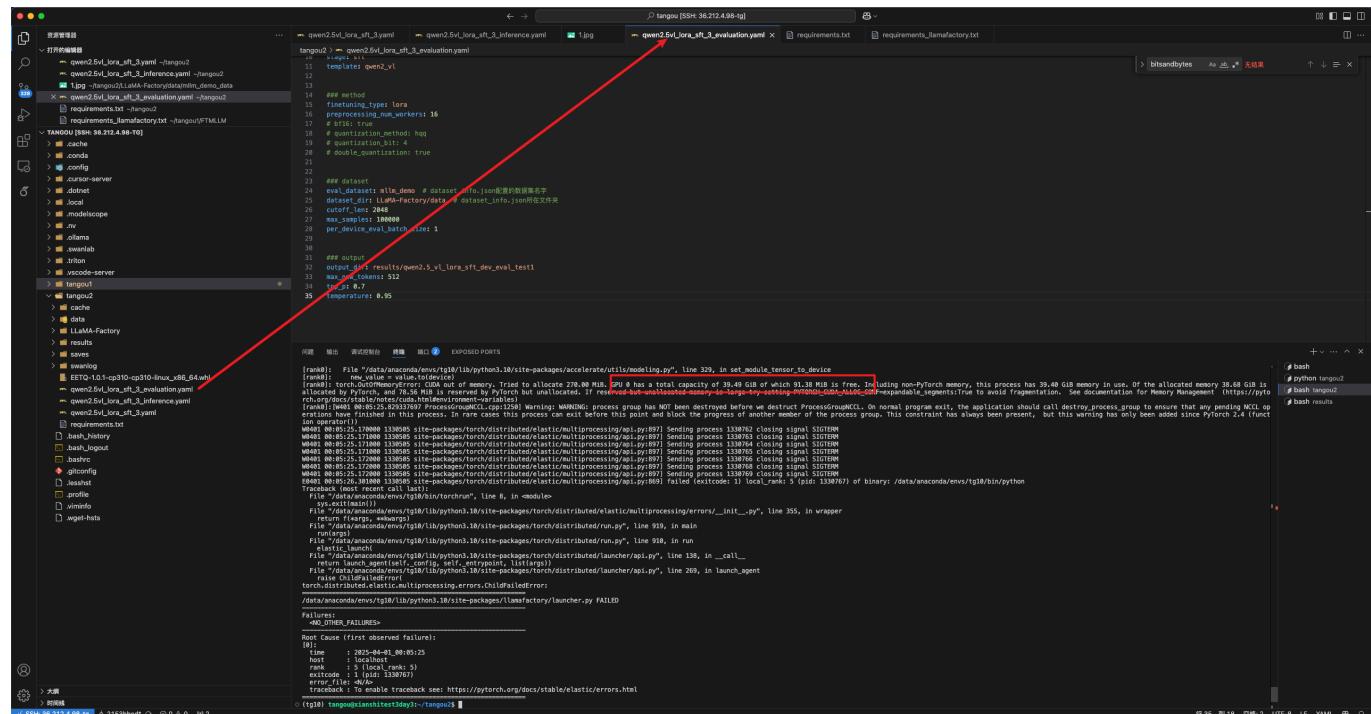


推理（这个图片在原本的模型上就一个训练过）



6. 评估 (实测: 32B评估时会爆显存)

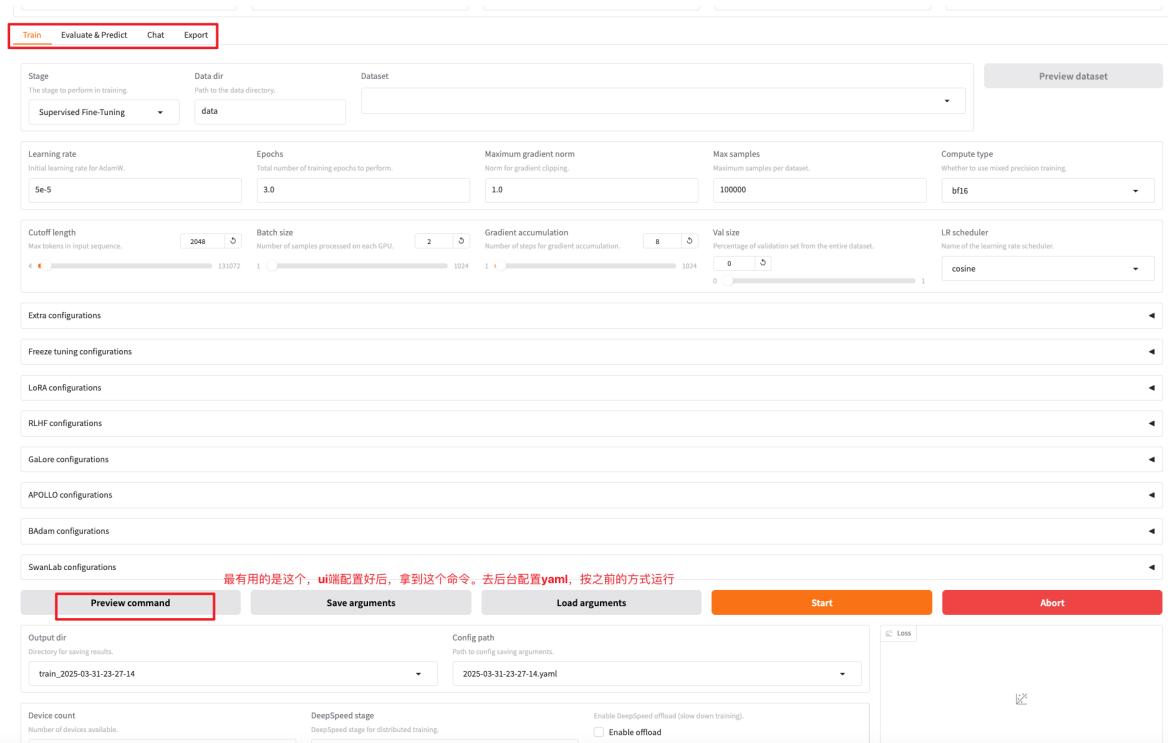
```
source /data/tools/setproxy.sh
conda activate tg10
NCCL_P2P_LEVEL=NVL HUGGINGFACE_HUB_CACHE="/data/huggingface/hub"
FORCE_TORCHRUN=1 CUDA_VISIBLE_DEVICES=0,1,2,3,4,5,6,7 llmfactory-client
train qwen2.5vl_lora_sft_3_evaluation.yaml
```



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
```

The screenshot shows the Hugging Face Model Card for the Qwen2.5-VL-7B-Instruct model. At the top, there's a navigation bar with links for Models, Datasets, Spaces, Posts, Docs, Enterprise, Pricing, and a user profile icon. A yellow banner at the top says "Hugging Face is way more fun with friends and colleagues! 😊 Join an organization" with a "Dismiss this message" button.

The main card area has a title "Qwen Qwen2.5-VL-7B-Instruct" with a "Follow" button. Below it are categories: Image-Text-to-Text, Transformers, Safetensors, English, qwen2_5_vl, multimodal, conversational, text-generation-inference, arxiv:2309.00071, arxiv:2409.12191, and arxiv:2308.12966. It also mentions "License: apache-2.0".

Below the categories are tabs for Model card, Files and versions, and Community. On the right, there are buttons for Edit model card, Train, Deploy, and Use this model.

The "Model card" section contains:

- Downloads last month:** 3,320,442 (with a line graph)
- Safetensors:** Model size: 8.29B params, Tensor type: BF16
- Inference Providers:** NEW, Hyperbolic (selected), Image-Text-to-Text, Examples

The "Inference Providers" section includes a text input field "Input a message to start chatting with Qwen/Qwen2.5-VL-7B-Instruct.", a "Your sentence here..." input field, a "Send" button, and "View Code" and "Maximize" buttons.