

- html:embed\_local\_images: true # 这个值设置是否将image转换成base64字符串toc: true # 这个值设置侧边栏
- 教程详细版
  - 服务器连接
  - 微调预备
  - 微调，这里用llamafactory提供的数据
  - 额外
  - export/merge model，合并然后评估（显存爆，和上面评估情况一样）
  - 量化评估
  - 测评，32B-4bit量化，7B全量。
    - 7B效果
    - 32B效果

---

**html: embed\_local\_images: true # 这个值设置是否将image转换成base64字符串  
toc: true # 这个值设置侧边栏**

---

## 教程详细版

---

视频版地址：链接: <https://pan.baidu.com/s/1GTNBNCrZ5hBw4w2CYj2jMeA?pwd=73c8>  
提取码: 73c8

vscode

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

扩展: 商店

chinese

The screenshot shows the Visual Studio Code Extensions Marketplace. A search bar at the top contains the text "chinese". Below the search bar, a list of extensions is displayed. The first extension in the list is "Chinese (Simplified) (简体中文) Language Pack for Visual Studio Code" by Microsoft, which is highlighted with a red border. Other extensions listed include "Chinese (Traditional) Language Pack for Visual Studio Code" by Microsoft, "ESLint Chinese Rules" by maggie, "Chinese Translation" by Compulim, "Chinese Lorem" by catlair, "Cline Chinese" by HybridTalentComputing, and "Chinese Lorem" by Traditional Chinese Lorem.

扩展名	名称	描述	作者	大小	评分	操作	
grid	Chinese (Simplified) (简体中文) Language Pack for Visual Studio Code	中文(简体)	Microsoft	在 SSH: 36.212.4.98-tg 中...	121K	5	安装
grid	Chinese (Traditional) Language Pack for Visual Studio Code	中文(繁體)	Microsoft				
ESLint	ESLint Chinese Rules	eslint中文规则提示插件	maggie		121K	5	安装
grid	Chinese Translation	Translates between Traditional and Simplified Chinese	Compulim		32K	5	安装
grid	Chinese Lorem	简体中文的乱数假文	catlair		44K		安装
grid	Cline Chinese	Cline中文汉化版, Cline是一款在您的 IDE 中运行的自主编程助手...	HybridTalentComputing		35K	5	安装
grid	Chinese Lorem	Traditional Chinese Lorem			39K	4	

扩展: 商店

remote

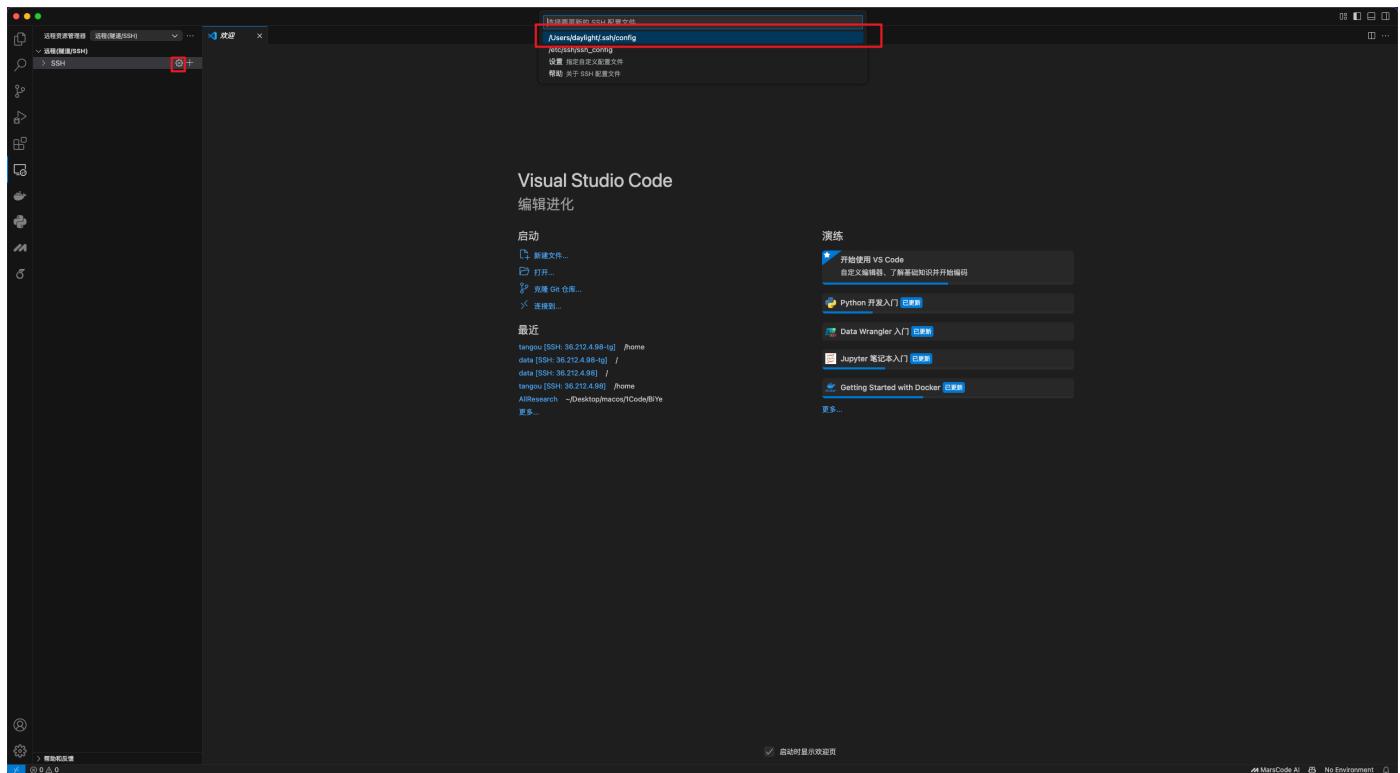
The screenshot shows the Visual Studio Code Extensions Marketplace. A search bar at the top contains the text "remote". Below the search bar, a list of extensions is displayed. The first extension in the list is "Remote - SSH" by Microsoft, which is highlighted with a red border. Other extensions listed include "Remote - SSH: Editing Configuration Files" by Microsoft, "Remote Explorer" by Microsoft, and "Gitpod Remote" by Gitpod.

扩展名	名称	描述	作者	延迟	操作
grid	Remote - SSH	Open any folder on a remote machine using SSH and take adva...	Microsoft	95ms	安装
grid	Remote - SSH: Editing Configuration Files	Edit SSH configuration files	Microsoft		
grid	Remote Explorer	View remote machines for SSH and Tunnels.	Microsoft	209ms	
grid	Gitpod Remote	Auto-installed within Classic workspaces, required to provision ...	Gitpod		

# 服务器连接

大家查看群文件自己的user和密码

vscode连接:

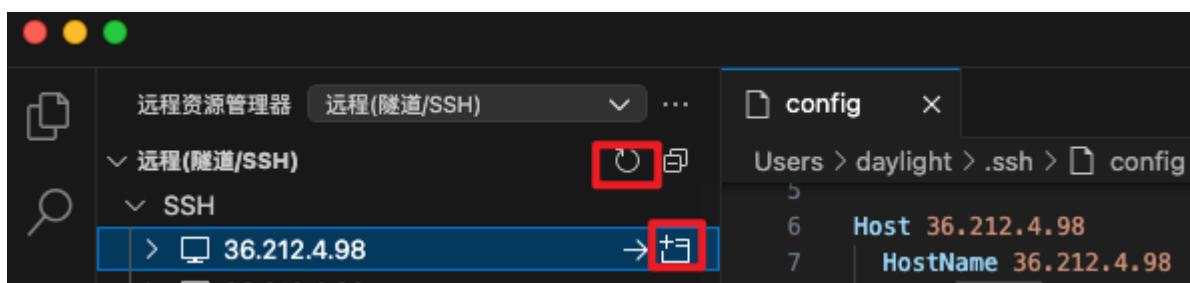


1. 复制到文件里面去

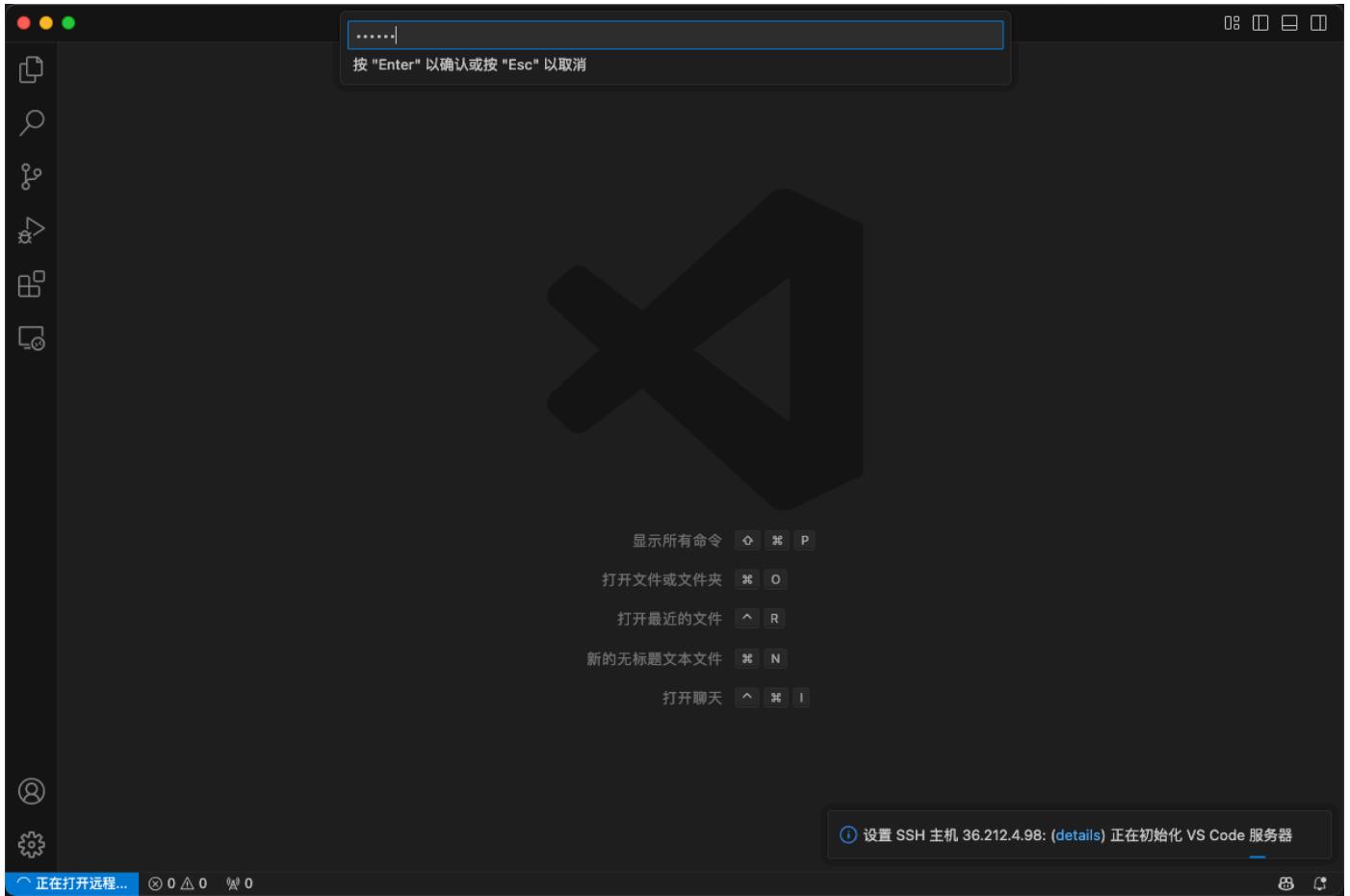
```
Host 36.212.4.98
HostName 36.212.4.98
User tangou
```

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

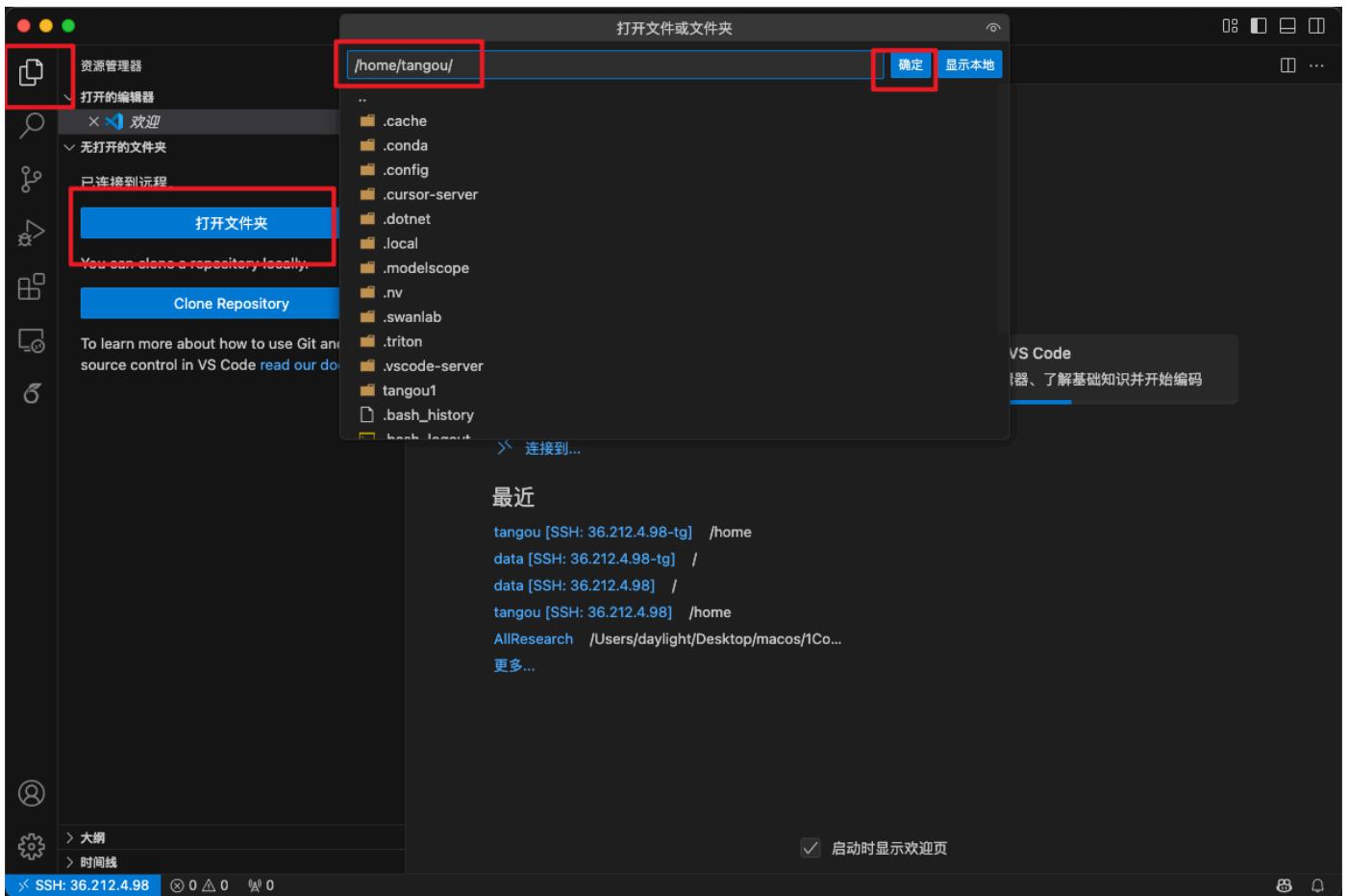
2. 点刷新，再打开文件



3. 输入密码123456，回车

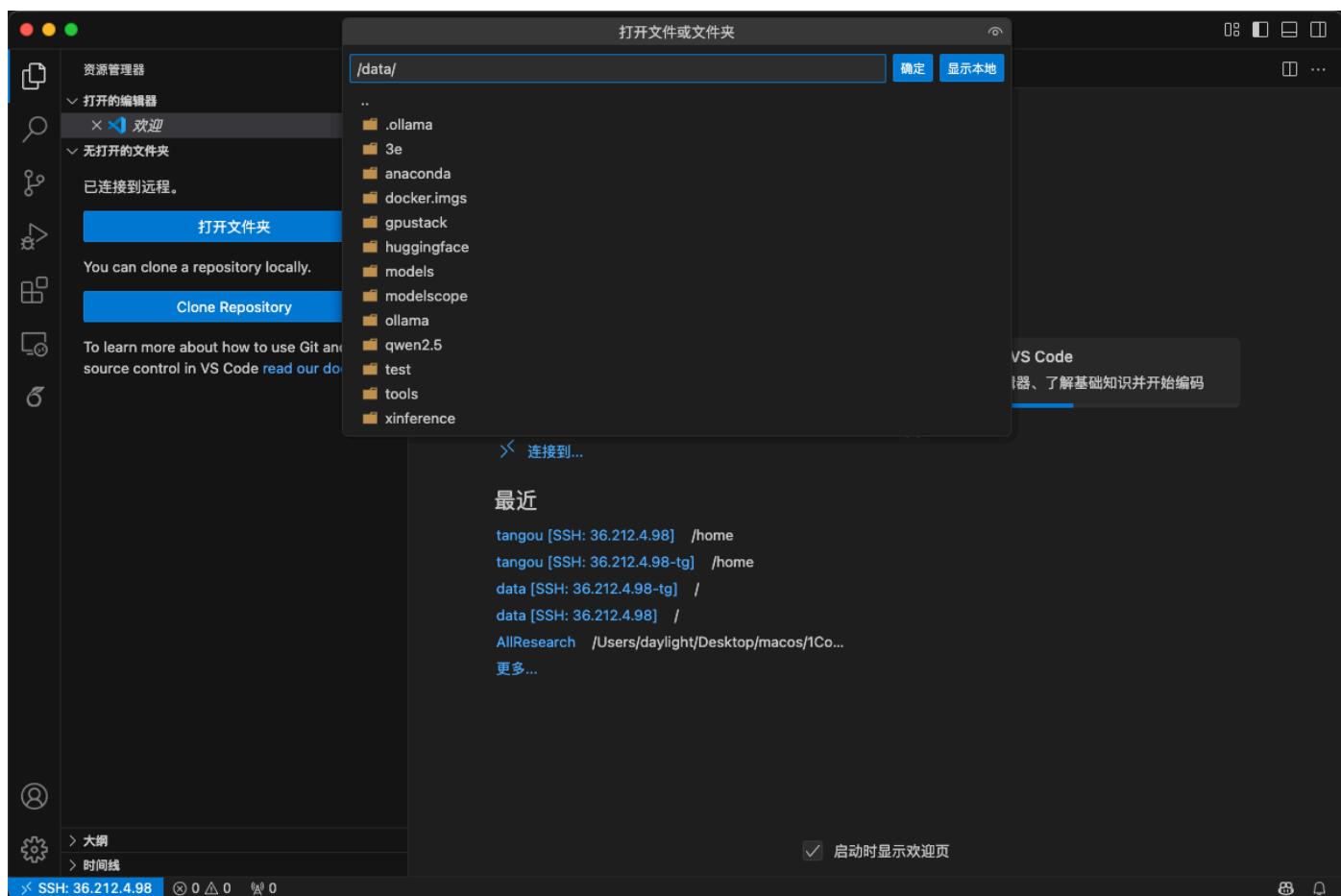


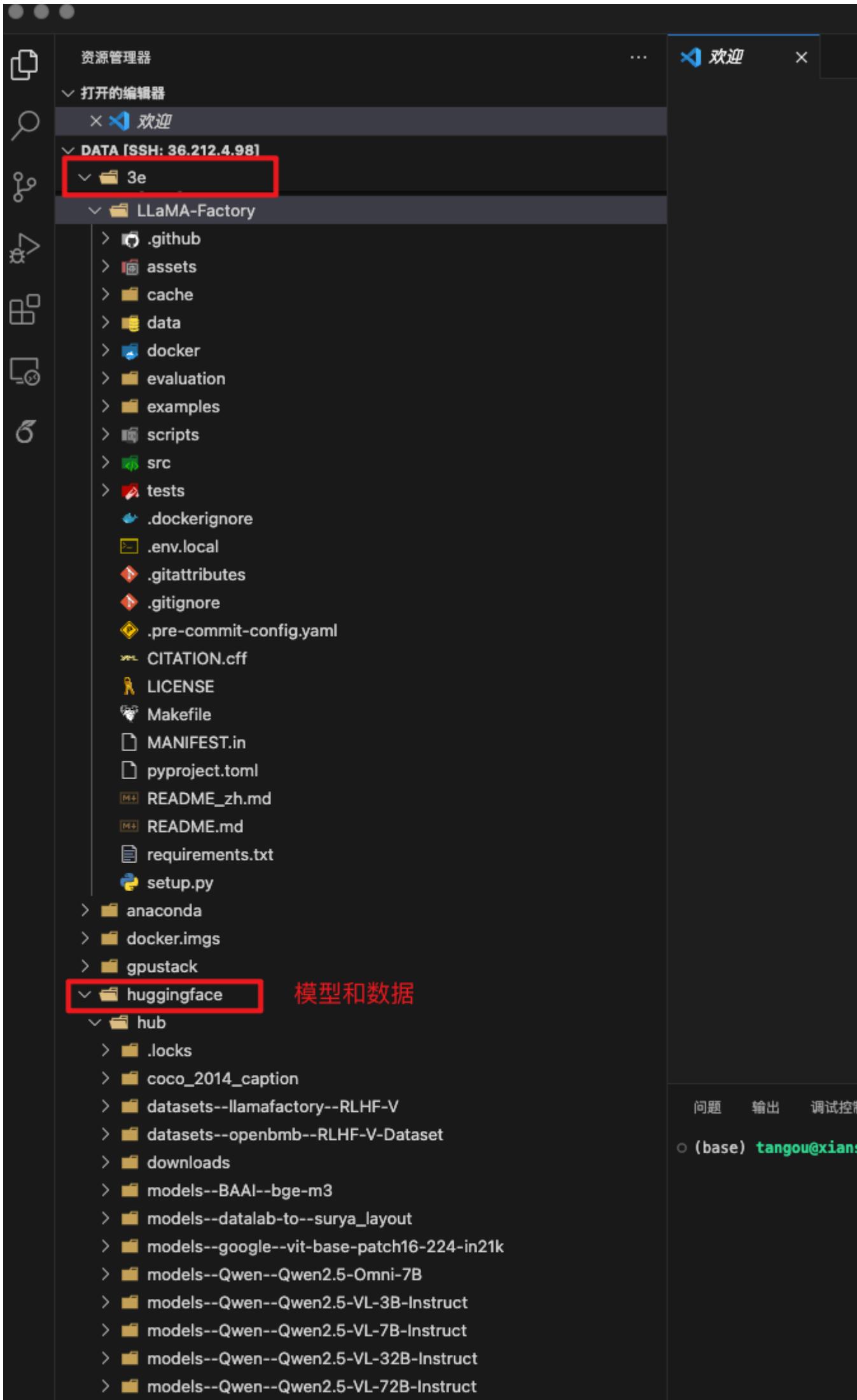
#### 4. 进入

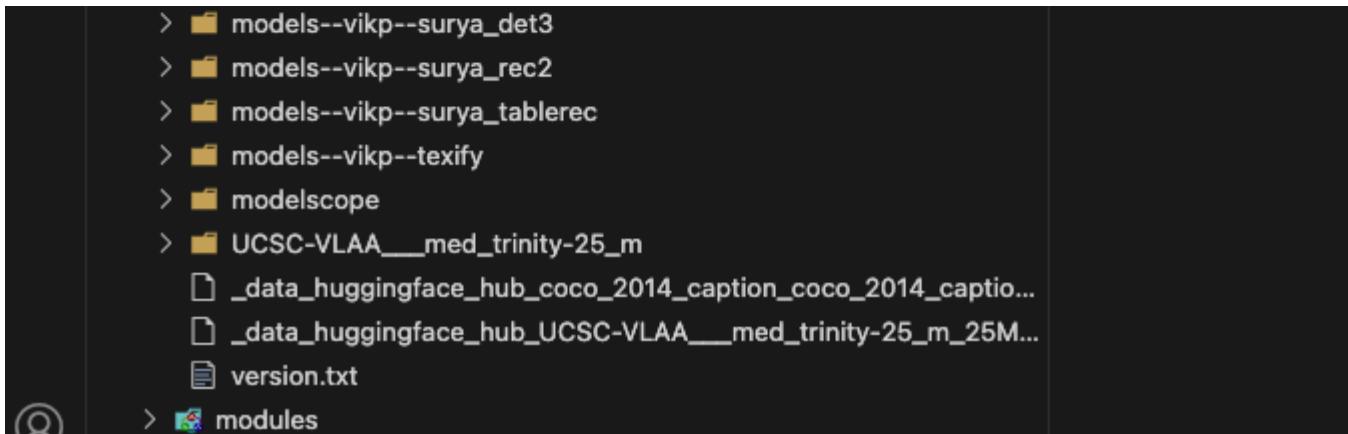




5. 回到第2步，再去开一个目录。这是共享模型数据的目录



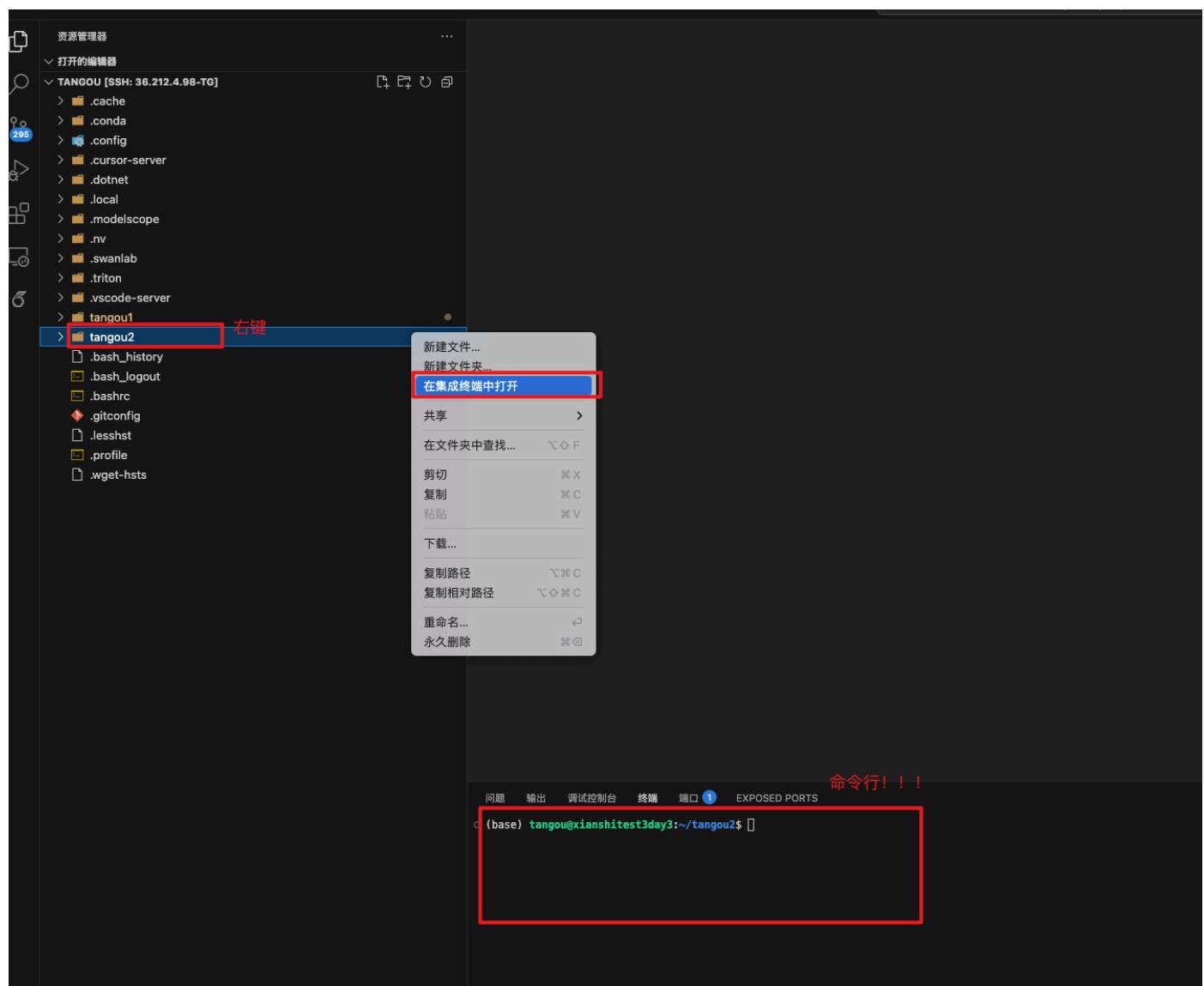




## 微调预备

- 环境变量：conda、ollama

### 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
```

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
```

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@xianshitest3day3:~/tangou2$ conda info --envs    查看conda有哪些环境
# conda environments:
#
base          * /data/anaconda
llama-factory      /data/anaconda/envs/llama-factory
alma           /data/anaconda/envs/alma
pptagent        /data/anaconda/envs/pptagent
qwen2.5-omni     /data/anaconda/envs/qwen2.5-omni
qwen2.5-vl      /data/anaconda/envs/qwen2.5-vl
vilm           /data/anaconda/envs/vilm

(base) tangou@xianshitest3day3:~/tangou2$ ollama list    ollama查看模型
NAME           ID          SIZE  MODIFIED
qwen2.5-vl-7B-Instruct:latest  8e7c8edcd38  14.2 GB  30 days ago
llama-3d:latest   7938764642687  1.2 GB  34 hours ago
bsahane/Qwen2.5-VL-7B-Instruct:Q4_K_M_benxh  dcccd488cacac  4.7 GB  2 days ago
qwen2.5:32b      9f13ba1299af  19 GB   4 days ago
omic-embed-text:latest  0a109f422b47  274 MB  7 days ago
llama-2d:vision:latest  08591fd1d725  71 MB   7 days ago
qwen2.5:72b      424baec1cf17  47 GB   8 days ago
Huzderu/deepspeak-r1-671b-2.5ibit:latest  bf800db59818  226 GB  12 days ago
deepspeak-r1-671b  739eb1b29ad7  404 GB  12 days ago

(base) tangou@xianshitest3day3:~/tangou2$ ollama run bsahane/Qwen2.5-VL-7B-Instruct:Q4_K_M_benxh    ollama运行其中一个模型
Hello! I'm here to help and answer 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!
>>> Send a message (/? for help)ctrl+C取消
```

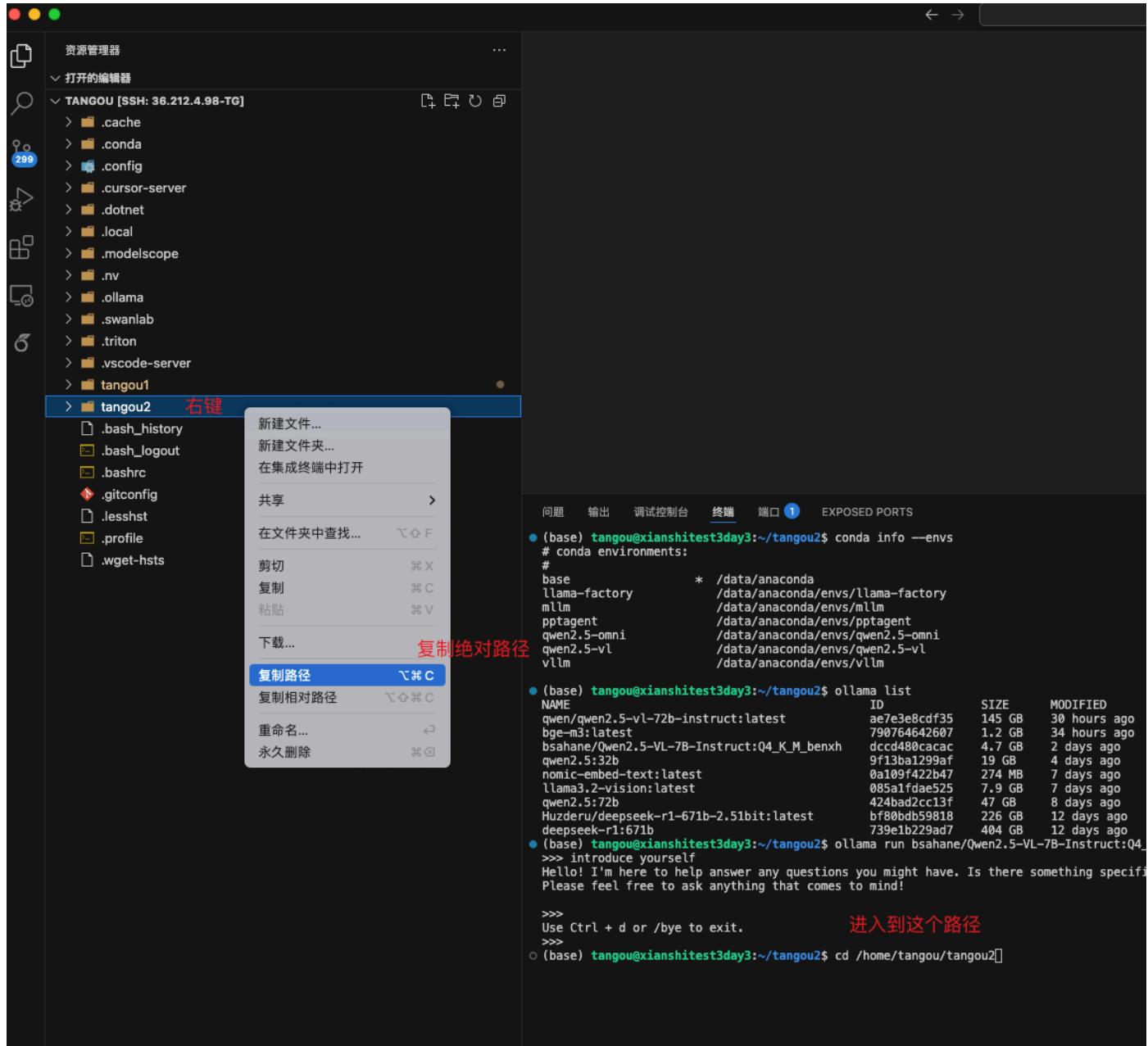
- vpn: 先不用管，知道这个就行

```
# 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

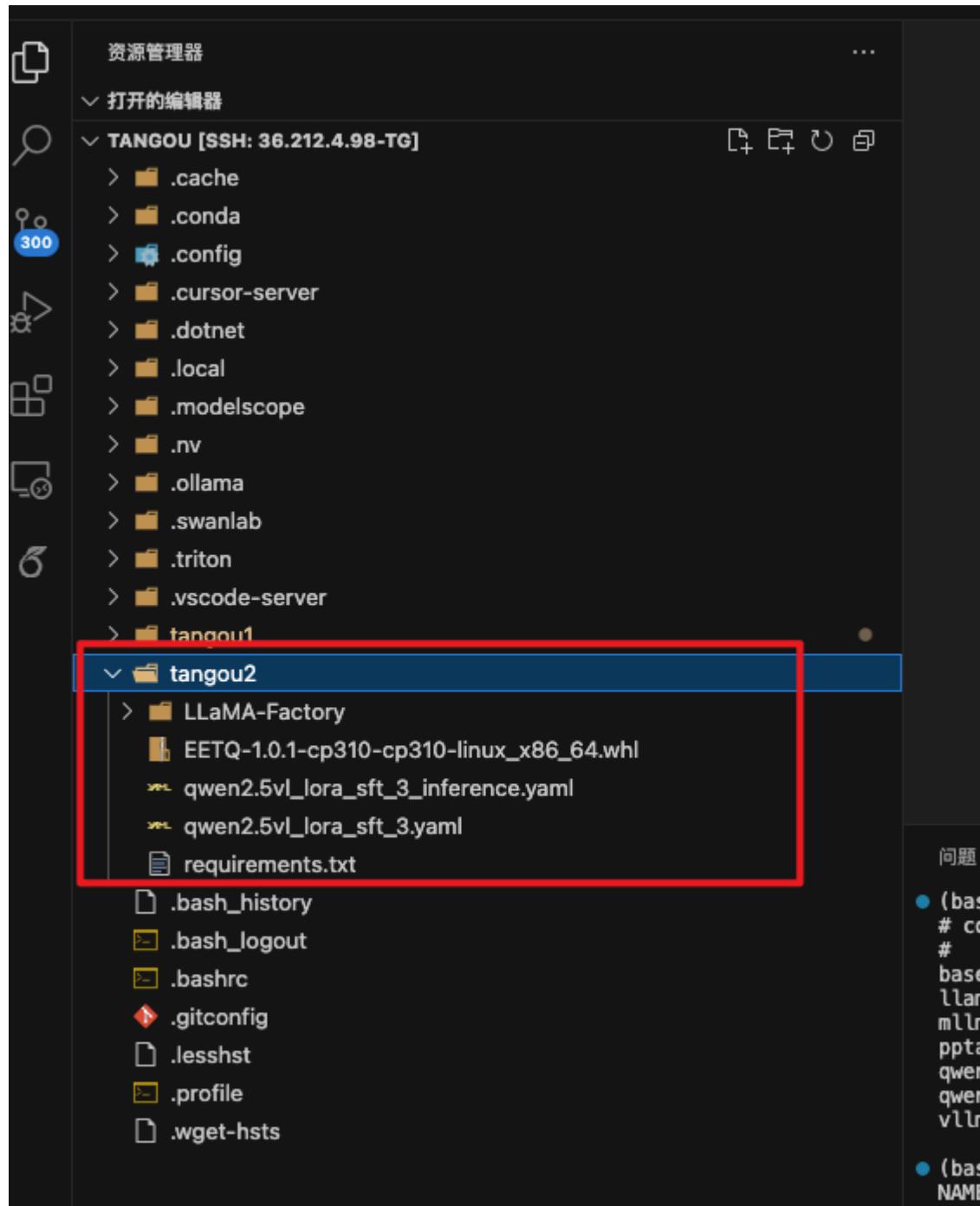
```
cd /home/tangou/tangou2 #你自己的路径
```



## 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
```

```

问题 输出 调试控制台 终端 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 ##

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-5.1-1_gnu
bz2                   pkgs/main/linux-64::bz2-1.0.8-h5ee18b_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-h6a678d5
libgcc-ng              pkgs/main/linux-64::libgcc-ng-10.2.1-h1234567_1
libomp                pkgs/main/linux-64::libomp-3.1.0-h1234567_1
libstdcxx-ng           pkgs/main/linux-64::libstdcxx-ng-11.2.0-h1234567_1
libuuid               pkgs/main/linux-64::libuuid-1.41.5-h5ee18b_0
ncurses               pkgs/main/linux-64::ncurses-6.4-h6a678d5_0
openssl               pkgs/main/linux-64::openssl-3.0.16-h5ee18b_0
pip                  pkgs/main/linux-64::pip-25.0-py310h6a4308_0
python                pkgs/main/linux-64::python-3.10.16-h70216_1
readline              pkgs/main/linux-64::readline-8.2-h5ee18b_0
setuptools            pkgs/main/linux-64::setuptools-59.5.0-py310h6a4308_0
sqlite               pkgs/main/linux-64::sqlite-3.45.3-h5ee18b_0
tk                   pkgs/main/linux-64::tk-8.6.14-h39e8969_0
tzdata               pkgs/main/nvcr://tzdata-2025a-h0ad1e81_0
wheel                pkgs/main/linux-64::wheel-0.45.1-py310h6a4308_0
xz                   pkgs/main/linux-64::xz-5.6.4-h5ee18b_1
zlib                 pkgs/main/linux-64::zlib-1.2.13-h5ee18b_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)

```



资源管理器

打开的编辑器

TANGOU [SSH: 36.212.4.98-TG]

- cache
- conda
- config
- cursor-server
- dotnet
- local
- modelscope
- nv
- ollama
- swanlab
- triton
- vscode-server
- tangou1
- tangou2
- LLaMA-Factory
- .github
- assets
- cache
- data
- belle\_multiturn
- hh\_rhf\_en
- mlm\_demo\_data
- 1.jpg
- 1.mp3
- 1.mp4
- 2.avi
- 2.jpg
- 2.wav
- 3.flac
- 3.jpg
- 3.mp4
- ultra\_chat
- alpaca\_en\_demo.json
- alpaca\_zh\_demo.json
- c4\_demo.json
- dataset\_info.json
- dpo\_en\_demo.json
- dpo\_zh\_demo.json
- glaive\_toolcall\_en\_demo.json
- glaive\_toolcall\_zh\_demo.json
- identity.json
- kto\_en\_demo.json
- mlm\_audio\_demo.json
- mlm\_demo.json
- mlm\_video\_demo.json
- README\_zh.md
- README.md
- wiki\_demo.txt
- docker

mlm\_demo.json

```

1 tangou@LLaMA-Factory:~/data> cat mlm_demo.json ...
2 {
3     "messages": [
4         {
5             "content": "<image>Who are they?", 
6             "role": "user"
7         },
8         {
9             "content": "They're Kane and Gretzka from Bayern Munich.", 
10            "role": "assistant"
11        },
12        {
13            "content": "What are they doing?<image>", 
14            "role": "user"
15        },
16        {
17            "content": "They are celebrating on the soccer field.", 
18            "role": "assistant"
19        }
20    ],
21    "images": [
22        "mlm_demo_data/1.jpg", 
23        "mlm_demo_data/1.jpg"
24    ]
25 },
26 {
27     "messages": [
28         {
29             "content": "<image>Who is he?", 
30             "role": "user"
31         },
32         {
33             "content": "He's Thomas Muller from Bayern Munich.", 
34             "role": "assistant"
35         },
36         {
37             "content": "Why is he on the ground?", 
38             "role": "user"
39         },
40         {
41             "content": "Because he's sliding on his knees to celebrate.", 
42             "role": "assistant"
43         }
44     ],
45     "images": [
46         "mlm_demo_data/2.jpg"
47     ]
48 }
49 
```

数据路径和左边文件夹对应，此为相对路径  
可配置为绝对路径

问题 输出 调试控制台 终端 端口 1 EXPOSED PORTS

(tg10) tangou@xianshitest3day3:~/tangou2\$ llmfactory-cli webui  
[2025-03-31 21:11:53,407] [INFO] [real\_accelerator.py:222:get\_accelerator] Setting ds\_accelerator to cuda (auto detect)  
\* 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

资源管理器

打开的编辑器

TANGOU [SSH: 36.212.4.98-TG]

- cache
- conda
- config
- cursor-server
- dotnet
- local
- modelscope
- nv
- ollama
- swanlab
- triton
- vscode-server
- tangou1
- tangou2
- LLaMA-Factory
- .github
- assets
- cache
- data
- belle\_multiturn
- hh\_rhf\_en
- mlm\_demo\_data
- 1.jpg
- 1.mp3
- 1.mp4
- 2.avi
- 2.jpg
- 2.wav
- 3.flac
- 3.jpg
- 3.mp4
- ultra\_chat
- alpaca\_en\_demo.json
- alpaca\_zh\_demo.json
- c4\_demo.json
- dataset\_info.json
- dpo\_en\_demo.json
- dpo\_zh\_demo.json
- glaive\_toolcall\_en\_demo.json
- glaive\_toolcall\_zh\_demo.json
- identity.json
- kto\_en\_demo.json
- mlm\_audio\_demo.json
- mlm\_demo.json
- mlm\_video\_demo.json
- README\_zh.md
- README.md
- wiki\_demo.txt
- docker

mlm\_demo.json

dataset\_info.json

```

1 tangou@LLaMA-Factory:~/data> cat dataset_info.json ...
2 {
3     "mlm_demo": {
4         "file_name": "mlm_demo.json", 
5         "formatting": "sharegpt",
6         "messages": "messages",
7         "images": "images"
8     },
9     "mlm_audio": {
10        "file_name": "mlm_audio_demo.json",
11        "formatting": "sharegpt",
12        "messages": "messages",
13        "audios": "audios"
14    },
15    "mlm_video": {
16        "file_name": "mlm_video_demo.json",
17        "formatting": "sharegpt",
18        "messages": "messages",
19        "videos": "videos"
20    },
21    "mlm_text": {
22        "file_name": "mlm_text_demo.json",
23        "formatting": "sharegpt",
24        "messages": "messages",
25        "text": "text"
26    },
27    "mlm_kto": {
28        "file_name": "mlm_kto_demo.json",
29        "formatting": "sharegpt",
30        "messages": "messages",
31        "kto": "kto"
32    }
33 } 
```

参照mlm\_demo.json问答对配置

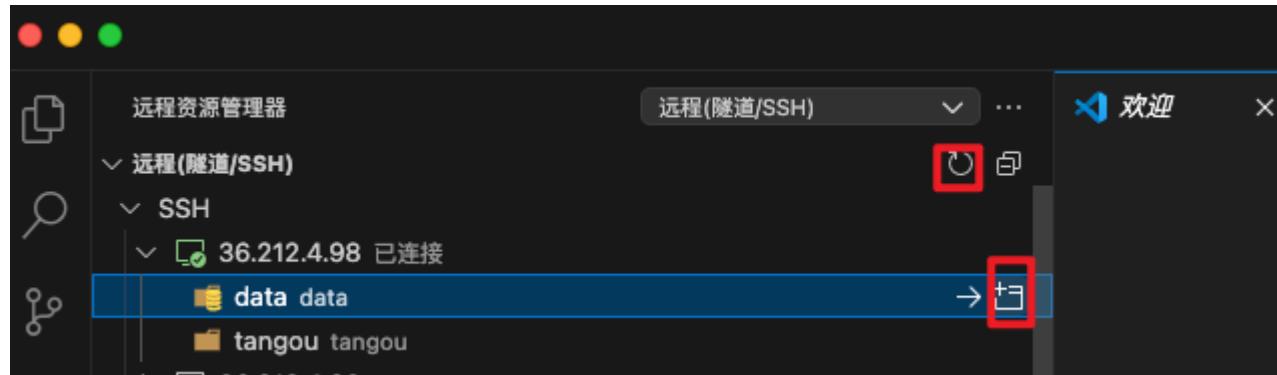
问题 输出 调试控制台 终端 端口 1 EXPOSED PORTS

(tg10) tangou@xianshitest3day3:~/tangou2\$ llmfactory-cli webui  
[2025-03-31 21:11:53,407] [INFO] [real\_accelerator.py:222:get\_accelerator] Setting ds\_accelerator to cuda (auto detect)  
\* Running on local URL: http://0.0.0.0:7860

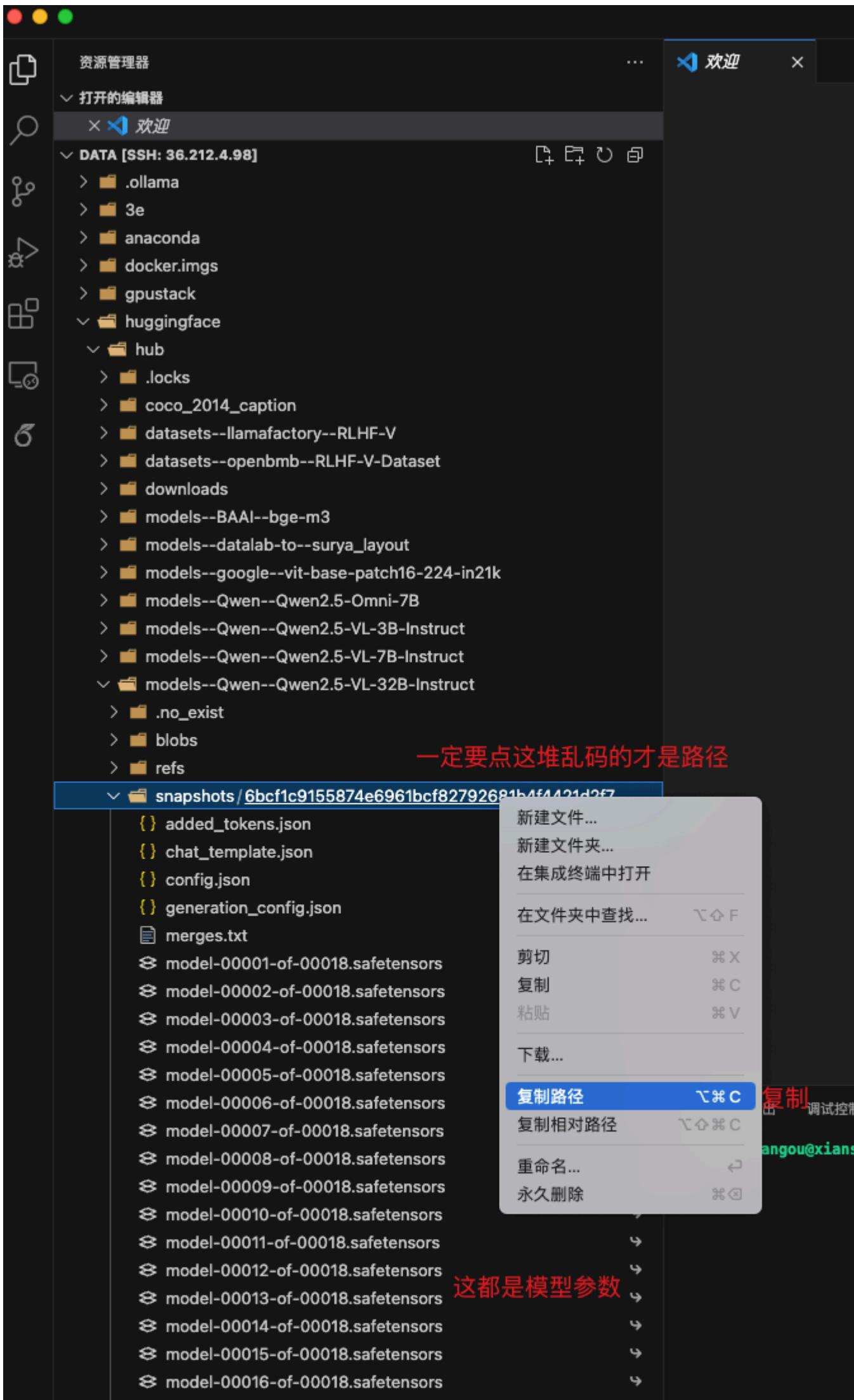
To create a public link, set `share=True` in `launch()`.

### 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



一定要点这堆乱码的才是路径

新建文件...

新建文件夹...

在集成终端中打开

在文件夹中查找... ⌘ F

剪切 ⌘ X

复制 ⌘ C

粘贴 ⌘ V

下载...

复制路径 ⌘ C

复制相对路径 ⌘ ⌘ C

重命名...

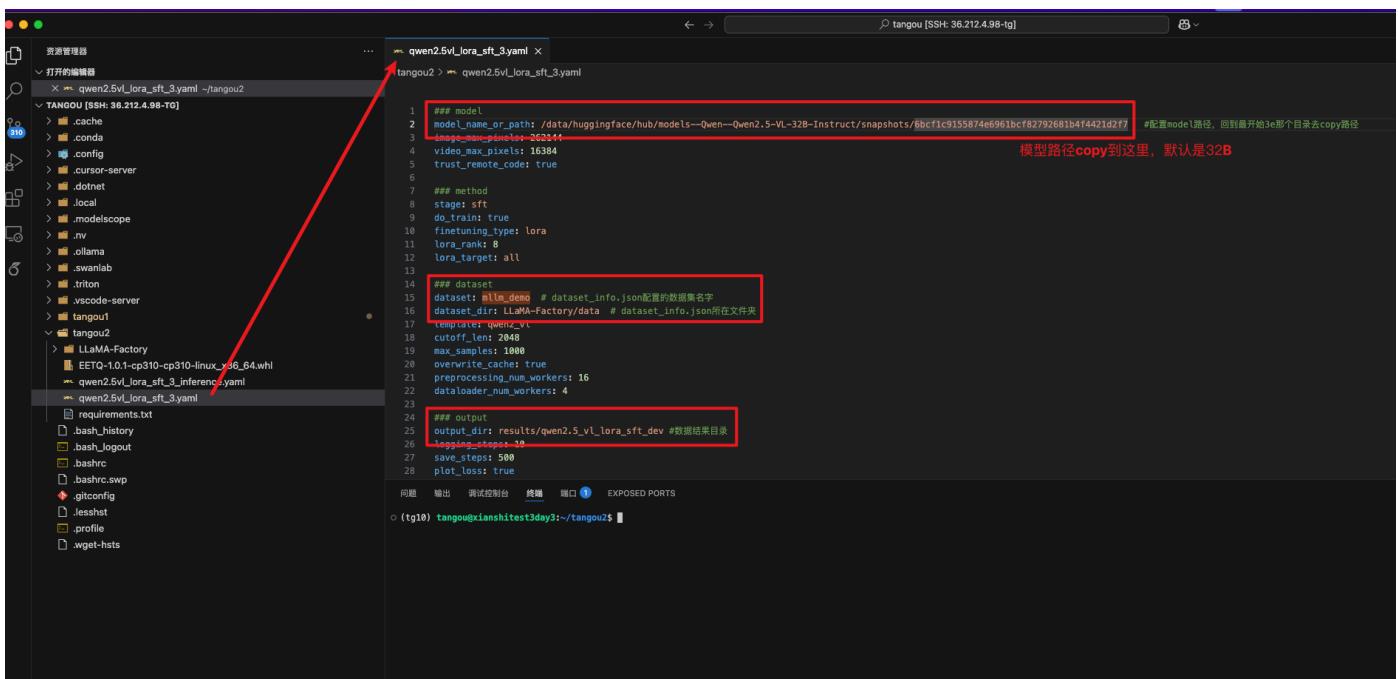
永久删除 ⌘ ⌘ D

这都是模型参数

↳ model-00017-of-00018.safetensors	↳
↳ model-00018-of-00018.safetensors	↳
{ } model.safetensors.index.json	↳
{ } preprocessor_config.json	↳
{ } special_tokens_map.json	↳
{ } tokenizer_config.json	↳
{ } tokenizer.json	↳
{ } vocab.json	↳

### 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



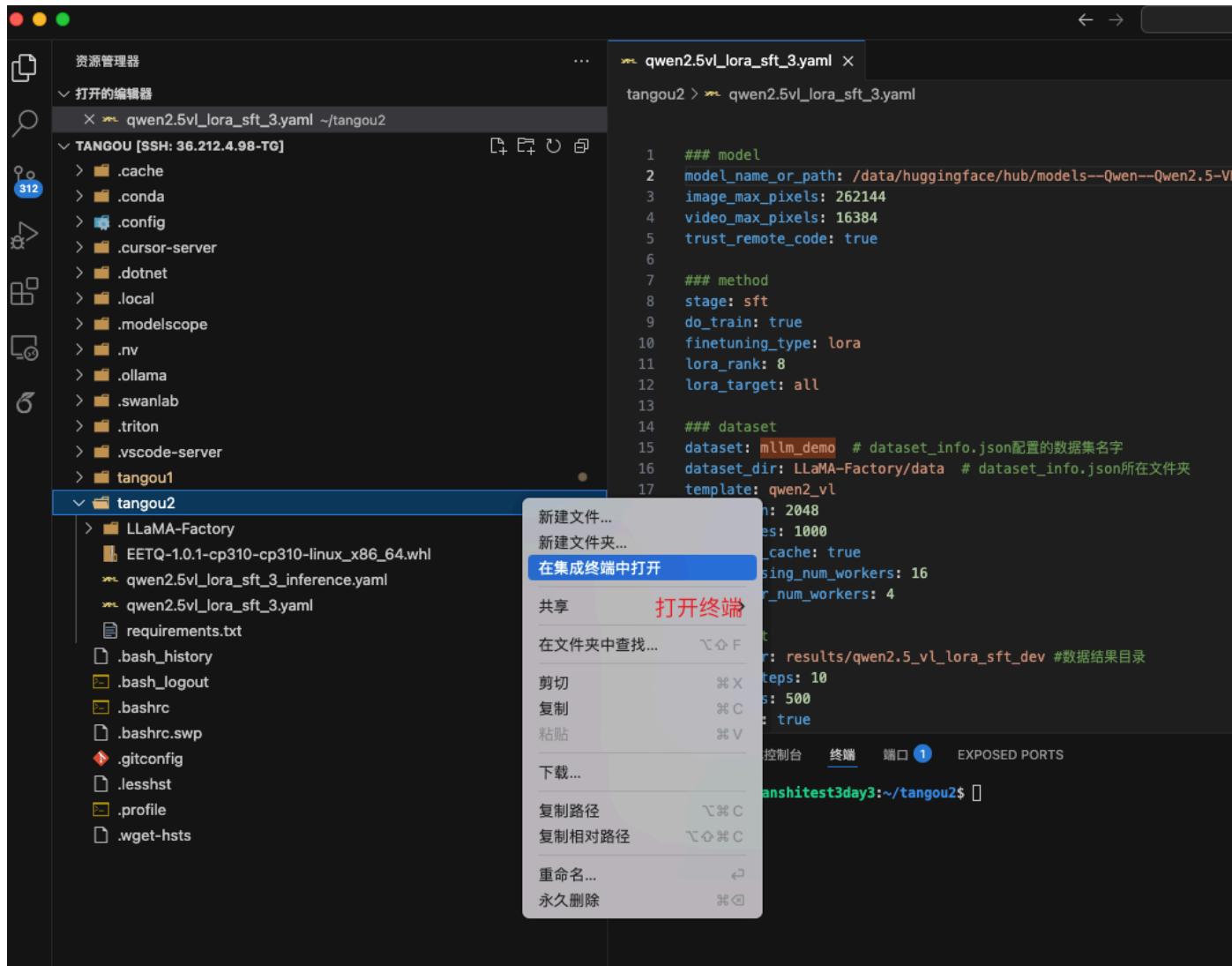
```

1 ##### model
2 model_name_or_path: /data/huggingface/hub/models--Owen--Qwen2.5-VL-32B-Instruct/snapshots/bbcf1c9155874eb961bcf8279268fb414421d7ff
3 max_seq_length: 2024
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: mlm_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 temp_folder: 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

```

### 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
```

资源管理器

```

qwen2.5vl_lora_sft_3.yaml
tangou2
TANGOU [SSH: 36.212.4.98-TG]
cache
conda
config
cursor-server
dotnet
local
modelscope
nv
ollama
swanlab
triton
vscode-server
tangou1
tangou2
LLM-A-Factory
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
bashrc.swp
gitconfig
Jeslist
profile
wget-hsts

```

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/b6c1c9155874e6961bcf82792681bf4421d2f #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: mlm_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 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

```

这里一定要切换成功

资源管理器

```

qwen2.5vl_lora_sft_3.yaml
minilm_demo.json
tangou2
TANGOU [SSH: 36.212.4.98-TG]
cache
conda
config
cursor-server
dotnet
local
modelscope
nv
ollama
swanlab
triton
vscode-server
tangou1
tangou2
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
gitconfig
Jeslist
profile
viminfo
wget-hsts

```

qwen2.5vl\_lora\_sft\_3.yaml

```

{} minilm_demo.json
tangou2 => qwen2.5vl_lora_sft_3.yaml
18 cutoff_len: 2048
19 max_samples: 1000
20 overwrite_cache: true

```

问题 提出 测试控制台 终端 网口 EXPOSED PORTS

```

(tg10) tangou@xianshitest3day3:~/tangou2$ source /data/tools/getproxy.sh
(tg10) tangou@xianshitest3day3:~/tangou2$ conda activate tg10
(tg10) tangou@xianshitest3day3:~/tangou2$ NCCL_P2P_LEVEL=NVL HUGGINGFACE_HUB_CACHE="--data/huggingface/hub" FORCE_TORCHRUN=1 CUDA_VISIBLE_DEVICES=4,5,6,7 llmfactory-cli train qwen2.5vl_lora_sft_3.yaml

```

运行进度条

资源管理器

```

.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

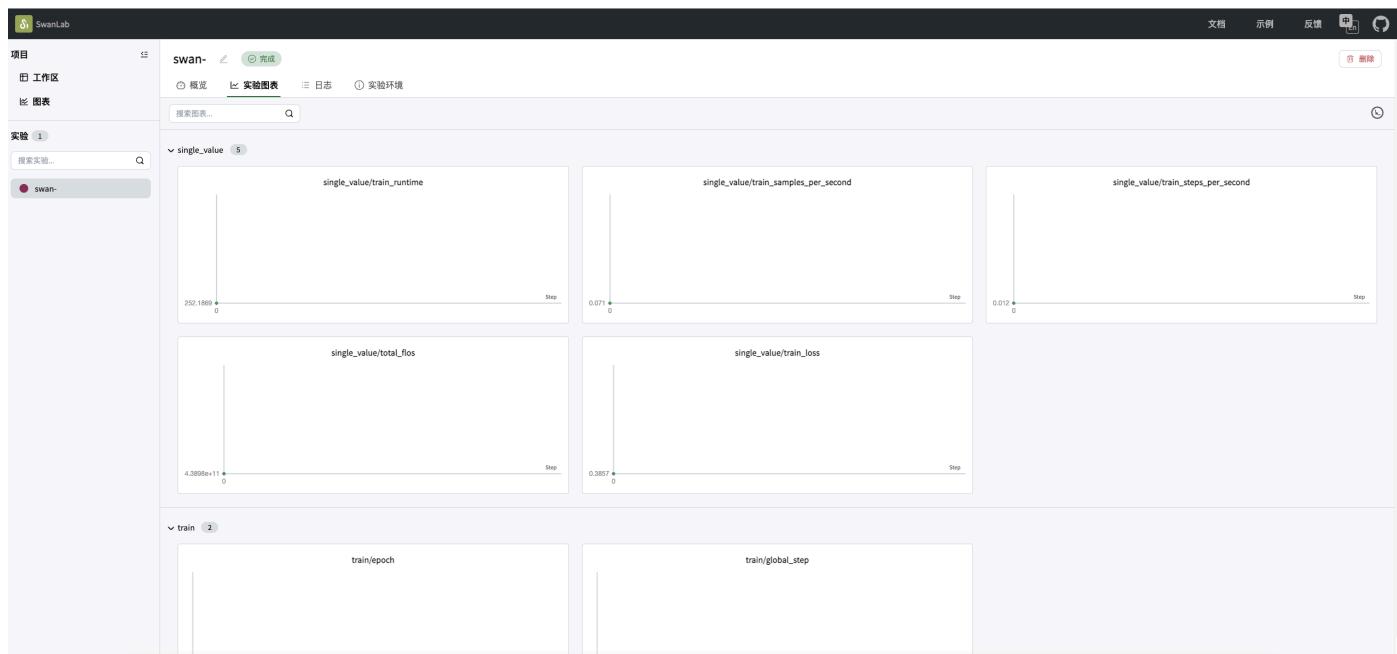
```

(tg10) tangou@xianshitest3day3:~/tangou2\$ swanlab watch swanlog  
swanboard: SwanLab Experiment Dashboard v0.4.12 ready in 15ms

+ Local: http://127.0.0.1:5092  
+ press **ctrl + c** to quit

问题	输出	调试控制台	终端	端口 2	EXPOSED PORTS	正
端口				转发地址		
o 5092					127.0.0.1:5092	
o 7860					localhost:7860	
<a href="#">添加端口</a>						
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 2023

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	34C	P0	37W / 250W	38505MiB / 40960MiB	0%	Default	Disabled	
1	NVIDIA A100-PCIE-40GB	Off	00000000:3E:00.0	Off			0	
N/A	34C	P0	38W / 250W	38505MiB / 40960MiB	0%	Default	Disabled	
2	NVIDIA A100-PCIE-40GB	Off	00000000:40:00.0	Off			0	
N/A	33C	P0	38W / 250W	38505MiB / 40960MiB	0%	Default	Disabled	
3	NVIDIA A100-PCIE-40GB	Off	00000000:41:00.0	Off			0	
N/A	35C	P0	40W / 250W	38505MiB / 40960MiB	0%	Default	Disabled	
4	NVIDIA A100-PCIE-40GB	Off	00000000:B1:00.0	Off			0	
N/A	34C	P0	36W / 250W	38505MiB / 40960MiB	0%	Default	Disabled	
5	NVIDIA A100-PCIE-40GB	Off	00000000:B2:00.0	Off			0	
N/A	35C	P0	41W / 250W	38505MiB / 40960MiB	0%	Default	Disabled	
6	NVIDIA A100-PCIE-40GB	Off	00000000:B4:00.0	Off			0	
N/A	34C	P0	36W / 250W	38505MiB / 40960MiB	0%	Default	Disabled	
7	NVIDIA A100-PCIE-40GB	Off	00000000:B5:00.0	Off			0	
N/A	34C	P0	37W / 250W	38505MiB / 40960MiB	0%	Default	Disabled	
<hr/>								
<b>Processes:</b>								
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 llmfactory-cli  
webchat qwen2.5vl_lora_sft_3_inference.yaml
```

The screenshot shows a terminal window with several files listed in the directory tree:

- qwen2.5vl\_lora\_sft\_3.yaml
- qwen2.5vl\_lora\_sft\_3\_inference.yaml
- tango2
- .cache
- .conda
- .config
- .cursor-server
- .dotnet
- .local
- .modelscope
- .nv
- .olama
- .swanlab
- .triton
- .vscode-server
- tangout
- 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
- .gitconfig
- .jessht
- .profile
- .viminfo
- wget-hsts

On the right side, the contents of `qwen2.5vl_lora_sft_3_inference.yaml` are displayed:

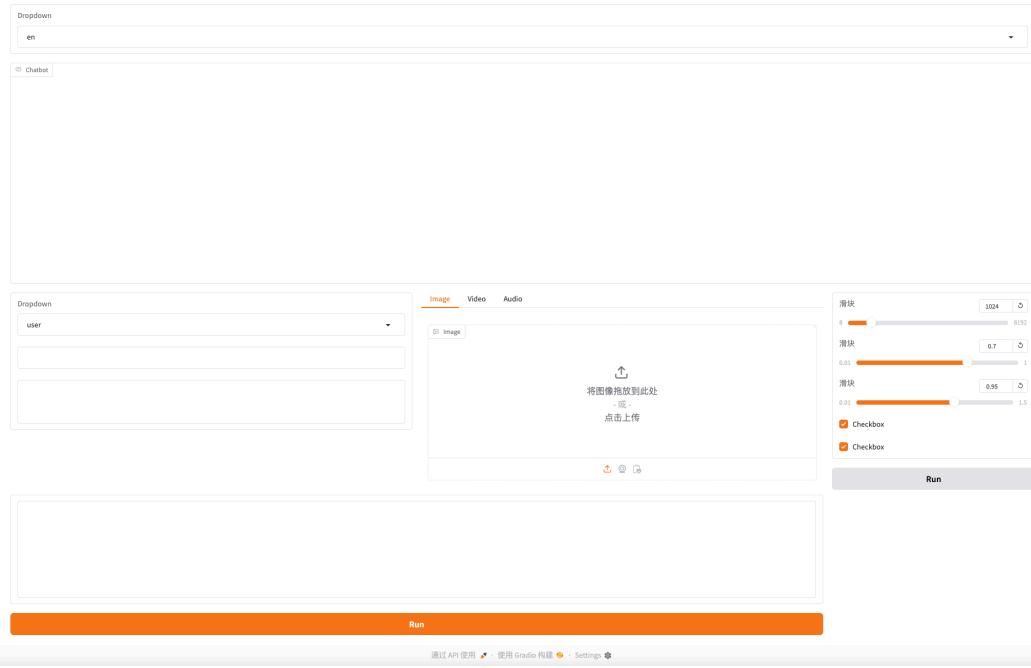
```
1 model_name_or_path: ./data/huggingface/hub/models--Owen--Qwen-2.5-VL-32B-Instruct/snapshots/6bcf1c9155874e6961bcf82792681b4f4421d2ff
2 adapter_name_or_path: ./models/qwen2.5_vl_lora_sft_dev
3 template: qwen2
4 infer_backend:vllm # choices: [huggingface, vlm]
5 trust_remote_code: true
```

The screenshot shows a terminal window with the "端口" tab selected, displaying the "EXPOSED PORTS" section.

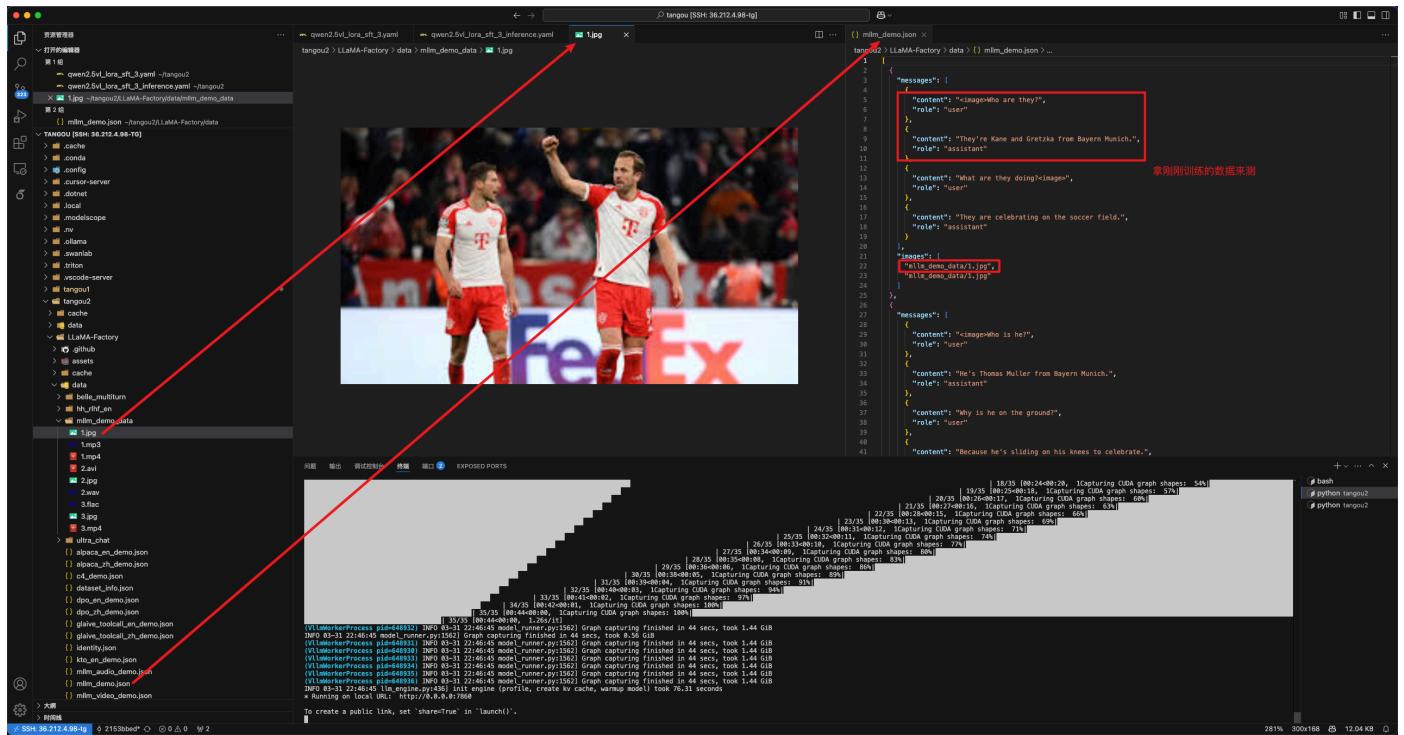
Output from the terminal:

```
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/it]
(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 0.56 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.py:436] 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()`.
```

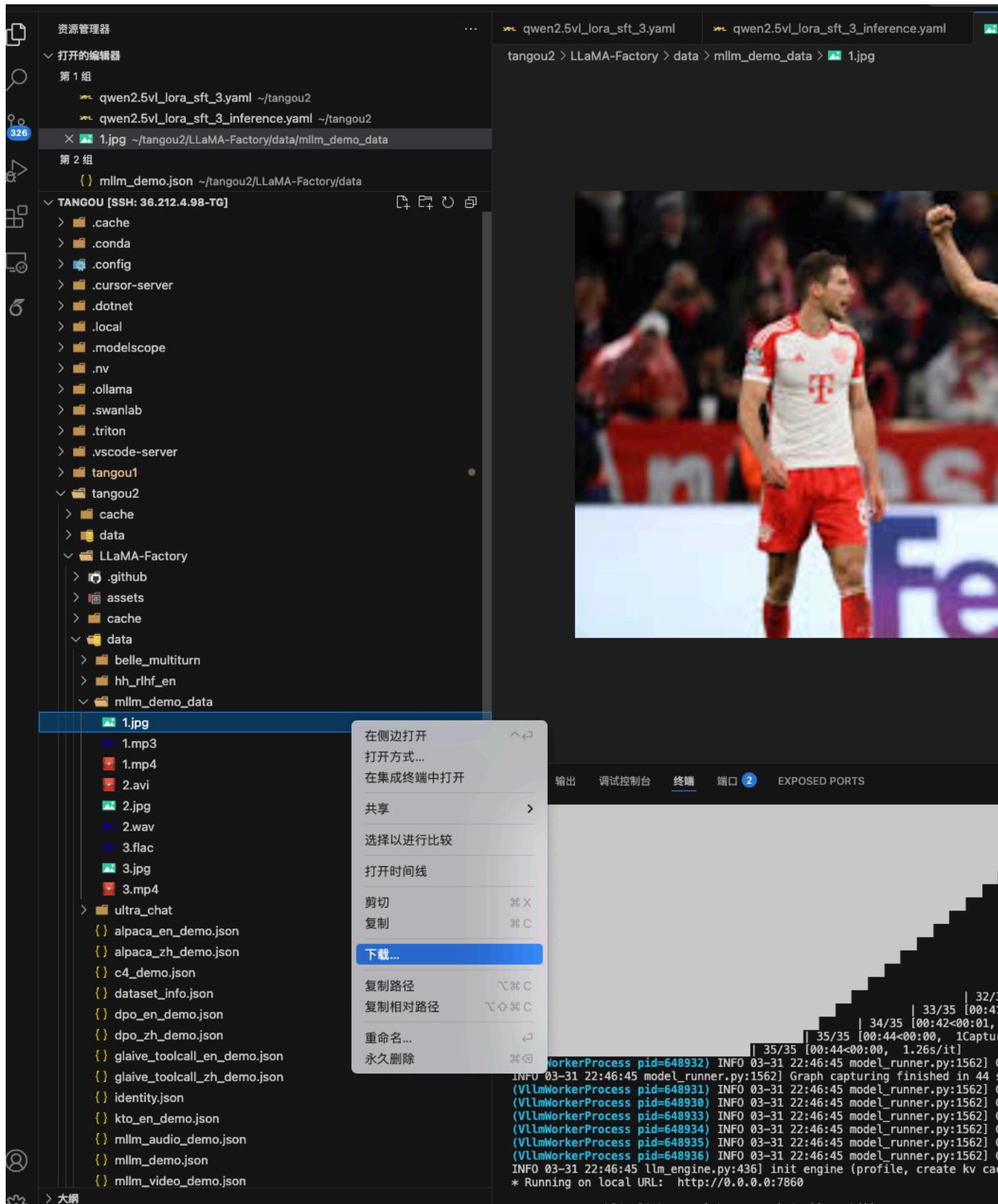
本地浏览器访问：<http://0.0.0.0:7860>



## 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）

Dropdown  
en

Chatbot

The image shows two football (soccer) players wearing red and white jerseys with the "FC Bayern München" logo, indicating they are players from Bayern Munich, a German professional football club.

- The player on the right is **Robert Lewandowski**, a Polish forward known for his prolific scoring record. He is widely recognized for his goal-scoring ability and has been a key player for Bayern Munich.
- The player on the left is **David Alaba**, an Austrian defender who has played a significant role for Bayern Munich and the Austrian national team.

Both players are celebrating, likely after scoring a goal or achieving a significant moment in the match. The FedEx logo in the background suggests this is during a game at the Allianz Arena, Bayern Munich's home stadium.

这个图片应该是被训练过，还会加粗名字

Dropdown  
user

Who are they?

Image   Video   Audio

Image

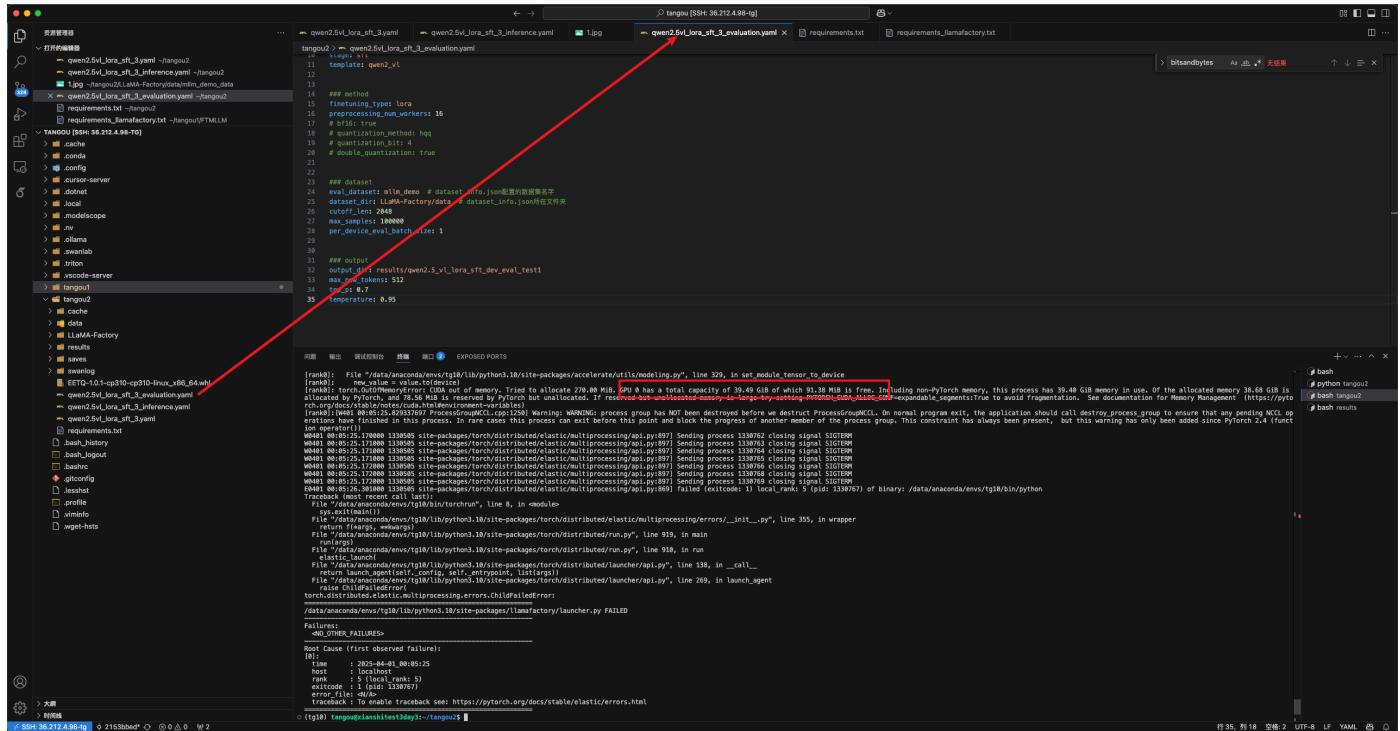
增块 0.01 增块 0.01 增块 0.01

Checkbox   Checkbox

Run

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
```



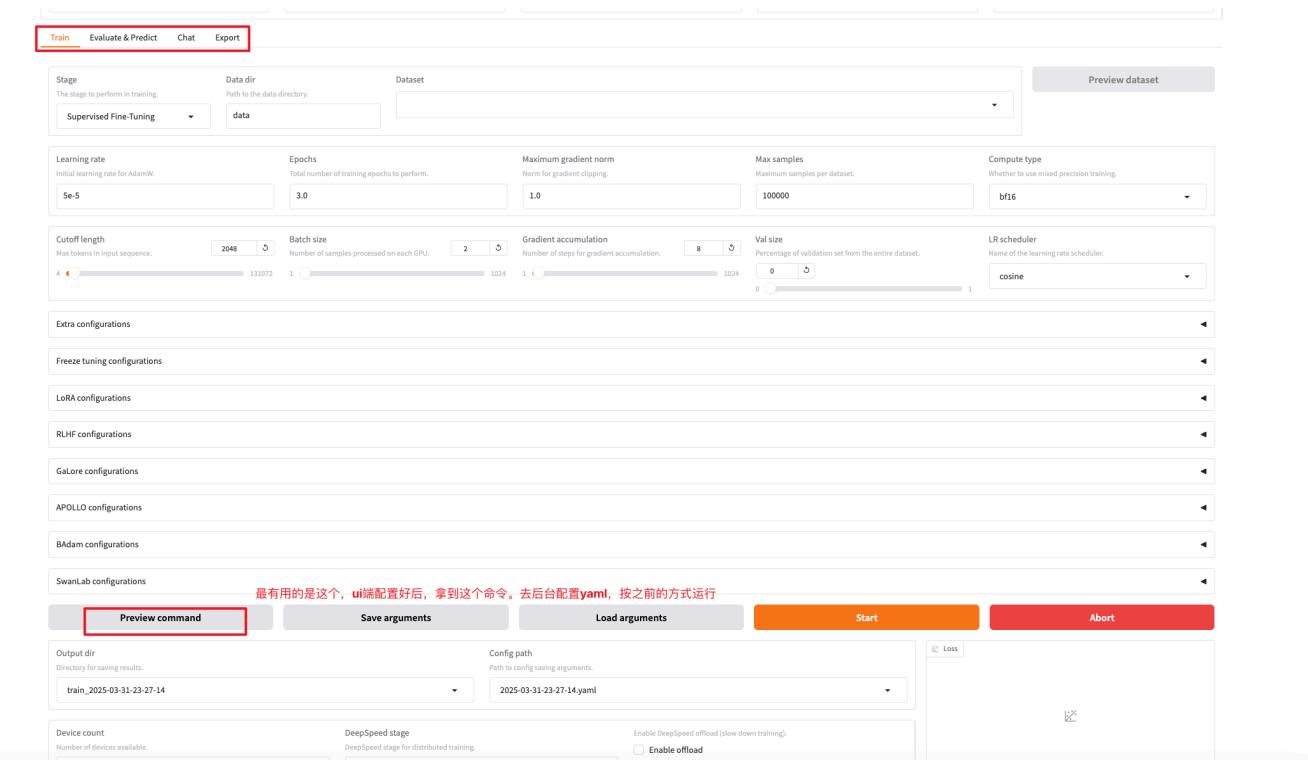
# 额外

## 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 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

License: apache-2.0

Model card Files and versions Community 38

Edit model card Train Deploy Use this model

## 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

# export/merge model, 合并然后评估（显存爆，和上面评估情况一样）

```
llamafactory-cli export qwen2.5vl_lora_sft_3_export.yaml
```

```
qwen2.5vl_lora_sft_3_export.yaml - tangou2
tangou2 > qwen2.5vl_lora_sft_3_export.yaml > README_tg.md
1  ### Note: DO NOT use quantized model or quantization_bit when merging lora adapters
2
3  ### model
4  model_name_or_path: /data/huggingface/hub/models--Qwen--Qwen2.5-VL-32B-Instruct/snapshots/6bc
5  adapter_name_or_path: results/qwen2.5_vl_lora_sft_dev
6  template: qwen2_vl
7  trust_remote_code: true
8
9  ### export
10 export_dir: results/export/qwen2.5_vl_lora_sft_dev
11 export_size: 5
12 export_device: cpu
13 export_legacy_format: false
```

# 导出完成

```
qwen2.5vl_lora_sft_3_export.yaml qwen2.5vl_lora_sft_3_evaluation.yaml README.md
tangou2 > qwen2.5vl_lora_sft_3_export.yaml
1  ### Note: DO NOT use quantized model or quantization_bit when merging lora adapters
2
3  ### model
4  model_name_or_path: /data/huggingface/hub/models--Qwen--Qwen2.5-VL-32B-Instruct/snapshots/6bcf1c9155874e6961bcf82792681b4f4421d2f7
5  adapter_name_or_path: results/qwen2.5_vl_lora_sft_dev
6  template: qwen2_vl
7  trust_remote_code: true
8
9  ### export
10 export_dir: results/export/qwen2.5_vl_lora_sft_dev
11 export_size: 5
12 export_device: cpu
13 export_legacy_format: false

问题 输出 调试控制台 端口 Exposed Ports
Loading checkpoint shards: 100% | 18/18
[0:04<00:00, 3.67it/s]
[INFO|modeling_utils.py:4970] 2025-04-21 00:57:39,463 >> All model checkpoint weights were used when initializing Qwen2.5_VLForConditionalGeneration.

[INFO|modeling_utils.py:4978] 2025-04-21 00:57:39,463 >> All the weights of Qwen2.5_VLForConditionalGeneration were initialized from the model checkpoint at /data/huggingface/hub/models--Qwen--Qwen2.5-VL-32B-Instruct/snapshots/6bcf1c9155874e6961bcf82792681b4f4421d2f7.
If your task is similar to the task the model of the checkpoint was trained on, you can already use Qwen2.5_VLForConditionalGeneration for predictions without further training.
[INFO|configuration_utils.py:1093] 2025-04-21 00:57:39,470 >> loading configuration file /data/huggingface/hub/models--Qwen--Qwen2.5-VL-32B-Instruct/snapshots/6bcf1c9155874e6961bcf82792681b4f4421d2f7/generation_config.json
[INFO|configuration_utils.py:1140] 2025-04-21 00:57:39,470 >> Generate config GenerationConfig {
    "bos_token_id": 151643,
    "do_sample": true,
    "eos_token_id": [
        151645,
        151645,
        151643
    ],
    "pad_token_id": 151643,
    "repetition_penalty": 1.05,
    "temperature": 1e-06
}

[INFO|2025-04-21 00:57:39] llmfactory.model.model_utils.attention:157 >> Using torch SDPA for faster training and inference.
[INFO|2025-04-21 00:58:43] llmfactory.model.adapter:157 >> Merged 1 adapter(s).
[INFO|2025-04-21 00:58:43] llmfactory.model.adapter:157 >> Loaded adapter(s): results/qwen2.5_vl_lora_sft_dev
[INFO|2025-04-21 00:58:43] llmfactory.model.loader:157 >> all params: 33,452,718,336
[INFO|2025-04-21 00:58:44] llmfactory.train.tuner:157 >> Convert model dtype to: torch.bfloat16.
[INFO|configuration_utils.py:423] 2025-04-21 00:58:44,009 >> Configuration saved in results/export/qwen2.5_vl_lora_sft_dev/config.json
[INFO|configuration_utils.py:909] 2025-04-21 00:58:44,010 >> Configuration saved in results/export/qwen2.5_vl_lora_sft_dev/generation_config.json
[INFO|modeling_utils.py:3048] 2025-04-21 01:00:16,483 >> The model is bigger than the maximum size per checkpoint (5GB) and is going to be split in 14 checkpoints across shards. You can find where each parameter has been saved in the index located at results/export/qwen2.5_vl_lora_sft_dev/model.safetensors.index.json.
[INFO|tokenization_utils_base.py:2500] 2025-04-21 01:00:16,485 >> tokenizer config file saved in results/export/qwen2.5_vl_lora_sft_dev/tokenizer_config.json
[INFO|tokenization_utils_base.py:2509] 2025-04-21 01:00:16,485 >> Special tokens file saved in results/export/qwen2.5_vl_lora_sft_dev/special_tokens_map.json
[INFO|image_processing_base.py:261] 2025-04-21 01:00:16,656 >> Image processor saved in results/export/qwen2.5_vl_lora_sft_dev/preprocessor_config.json
[INFO|tokenization_utils_base.py:2500] 2025-04-21 01:00:16,699 >> tokenizer config file saved in results/export/qwen2.5_vl_lora_sft_dev/tokenizer_config.json
[INFO|tokenization_utils_base.py:2509] 2025-04-21 01:00:16,700 >> Special tokens file saved in results/export/qwen2.5_vl_lora_sft_dev/special_tokens_map.json
[INFO|processing_utils.py:638] 2025-04-21 01:00:17,560 >> chat template saved in results/export/qwen2.5_vl_lora_sft_dev/chat_template.json
[INFO|2025-04-21 01:00:17] llmfactory.train.tuner:157 >> Ollama modelfile saved in results/export/qwen2.5_vl_lora_sft_dev/Modelfile
o (tg10) tangou@xianshitest3day3:~/tangou2$
```

```
ollama create qwen2.5_vl_custom -f
/home/tangou/tangou2/results/export/qwen2.5_vl_lora_sft_dev/Modelfile
ollama run qwen2.5_vl_custom
```

## 评估

```
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-cli train
qwen2.5vl_lora_sft_3_evaluation_no_lora.yaml
```

## 爆显存

```
[rank0]: File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/transformers/models/auto/auto_factory.py", line 564, in from_pretrained
[rank0]:     return model_class.from_pretrained(
[rank0]:     File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/transformers/modeling_utils.py", line 262, in _wrapper
[rank0]:     return func(*args, **kwargs)
[rank0]:     File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/transformers/modeling_utils.py", line 4319, in from_pretrained
[rank0]:         ) = cls._load_pretrained_model(
[rank0]:         File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/transformers/modeling_utils.py", line 4897, in _load_pretrained_model
[rank0]:             new_error_msgs, offload_index, state_dict_index = _load_state_dict_into_meta_model(
[rank0]:             File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/transformers/modeling_utils.py", line 896, in _load_state_dict_into_meta_model
[rank0]:                 set_module_tensor_to_device(model, param_name, param_device, **set_module_kwargs)
[rank0]:                 File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/accelerate/utils/modeling.py", line 329, in set_module_tensor_to_device
[rank0]:                     new_value = value.to(device)
[rank0]: torch.cuda.OutOfMemoryError: CUDA out of memory. Tried to allocate 270.00 MiB. GPU 0 has a total capacity of 39.49 GiB of which 115.38 MiB is free. Including non-PyTorch memory, this process has 39.37 GiB memory in use. Of the allocated memory 38.68 GiB is allocated by PyTorch, and 78.56 MiB is reserved by PyTorch but unallocated. If reserved but unallocated memory is large try setting PYTORCH_CUDA_ALLOC_CONF=expandable_segments=True to avoid fragmentation. See documentation for Memory Management (https://pytorch.org/docs/stable/notes/cuda.html#environment-variables)
[rank0]: [W421 01:13:30.390980251 ProcessGroupNCCL.cpp:1250] Warning: WARNING: process group has NOT been destroyed before we destruct ProcessGroupNCCL. On normal program exit, the application should call destroy_process_group to ensure that any pending NCCL operations have finished in this process. In rare cases this process can exit before this point and block the progress of another member of the process group. This constraint has always been present, but this warning has only been added since PyTorch 2.4 (function operator())
Loading checkpoint shards: 50%|██████████| 7/14 [00:24<00:23, 3.36s/it]W0421 01:13:32.23800
0 2732395 site-packages/torch/distributed/elastic/multiprocessing/api.py:897] Sending process 2732613 closing signal SIGTERM
W0421 01:13:32.239000 2732395 site-packages/torch/distributed/elastic/multiprocessing/api.py:897] Sending process 2732614 closing signal SIGTERM
W0421 01:13:32.239000 2732395 site-packages/torch/distributed/elastic/multiprocessing/api.py:897] Sending process 2732615 closing signal SIGTERM
W0421 01:13:32.240000 2732395 site-packages/torch/distributed/elastic/multiprocessing/api.py:897] Sending process 2732616 closing signal SIGTERM
W0421 01:13:32.240000 2732395 site-packages/torch/distributed/elastic/multiprocessing/api.py:897] Sending process 2732617 closing signal SIGTERM
W0421 01:13:32.240000 2732395 site-packages/torch/distributed/elastic/multiprocessing/api.py:897] Sending process 2732618 closing signal SIGTERM
W0421 01:13:32.241000 2732395 site-packages/torch/distributed/elastic/multiprocessing/api.py:897] Sending process 2732619 closing signal SIGTERM
E0421 01:13:33.770000 2732395 site-packages/torch/distributed/elastic/multiprocessing/api.py:869] failed (exitcode: 1) local_rank: 0 (pid: 2732612) of binary: /data/anaconda/envs/tg10/bin/python
Traceback (most recent call last):
  File "/data/anaconda/envs/tg10/bin/torchrun", line 8, in <module>
    sys.exit(main())
  File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/torch/distributed/elastic/multiprocessing/errors/__init__.py", line 355, in wrapper
    return f(*args, **kwargs)
  File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/torch/distributed/run.py", line 919, in main
    run(args)
  File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/torch/distributed/run.py", line 910, in run
    elastic_launch(
  File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/torch/distributed/launcher/api.py", line 138, in __call__
    return launch_agent(self._config, self._entrypoint, list(args))
  File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/torch/distributed/launcher/api.py", line 269, in launch_agent
    raise ChildFailedError(
torch.distributed.elastic.multiprocessing.errors.ChildFailedError:
Failures:
<NO_OTHER_FAILURES>
Root Cause (first observed failure):
[0]:
  time      : 2025-04-21_01:13:32
  host      : localhost
  rank      : 0 (local_rank: 0)
  exitcode   : 1 (pid: 2732612)
  error_file: </A>
  traceback : To enable traceback see: https://pytorch.org/docs/stable/elastic/errors.html
```

## 量化评估

32B模型微调后，8bit量化爆显存，8卡每张卡占40G左右，4bit量化每张卡占20G左右，如下结果

修改qwen2.5vl\_lora\_sft\_3\_evaluation.yaml的量化，bit，为4和8，如果不需要就把其全部注释

```
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-cli train
qwen2.5vl_lora_sft_3_evaluation.yaml
```

资源管理器

TANGOU [SSH: 36.212.4.98]

```

> .modelscope
> .nv
> .ollama
> .rest-client
> .swanlab
> .trae-server
> .triton
> .vscode-server
> tangou1
< tangou2
> cache
> data
> LLaMA-Factory
< results
> export
> qwen2.5_vl_lora_sft_dev
< qwen2.5_vl_lora_sft_dev_eval_test1
  {} all_results.json
  {} generated_predictions.json
  {} predict_results.json
> results_new
> swanlog
> build01.log
> EETQ-1.0.1-cp310-cp310-linux_x86_64.whl
> qwen2.5vl_lora_sft_3_evaluation_no_lora.yaml
> qwen2.5vl_lora_sft_3_evaluation.yaml
> qwen2.5vl_lora_sft_3_export.yaml
> qwen2.5vl_lora_sft_3_inference.yaml
> qwen2.5vl_lora_sft_3.yaml
requirements.txt
test-01.yaml
test-01.yaml-
test-02-infer.yaml
test-02.yaml
test-03-dis-eval.yaml
test-03-new.yaml

```

qwen2.5vl\_lora\_sft\_3\_evaluation.yaml

```

1  ### model
2  model_name_or_path: /data/huggingface/hub/models--Qwen--Qwen2.5-VL-32B-Instruct/snap
3  copy路径
4  adapter_name_or_path: results/qwen2.5_vl_lora_sft_dev
5  trust_remote_code: true
6  do_predict: true
7  predict_with_generate: true
8  flash_attn: auto
9
10 stage: sft
11 template: qwen2_vl
12
13
14 ### method
15 finetuning_type: lora
16 preprocessing_num_workers: 16
17
18 ### quantization
19 quantization_method: bitsandbytes
20 quantization_bit: 4
21 double_quantization: true
22
23
24 ### dataset
25 eval_dataset: mllm_demo # dataset_info.json配置的数据集名字
26 dataset_dir: LLaMA-Factory/data # dataset_info.json所在文件夹
27 cutoff_len: 2048
28 max_samples: 100000
29 per_device_eval_batch_size: 1
30
31
32 ### output
33 output_dir: results/qwen2.5_vl_lora_sft_dev_eval_test1
34 max_new_tokens: 512

```

问题 输出 调试控制台 端口 1 Exposed Ports

Prefix dict has been built successfully.  
Dump cache file failed.  
Traceback (most recent call last):  
File "/data/anaconda/envs/tg10/lib/python3.10/site-packages/jieba/\_\_init\_\_.py", line 154, in  
\_replace\_file(fp, cache\_file)  
PermissionError: [Errno 1] Operation not permitted: '/tmp/tmppp2\_27f8d' -> '/tmp/jieba.cache'  
Loading model cost 0.577 seconds.  
Prefix dict has been built successfully.  
Dump cache file failed.

(base) cangou@x280:~/Desktop\$

Mon Apr 21 01:21:41 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	35C	P0	64W / 250W	24796MiB / 40960MiB	100%	Default	Disabled	
1	NVIDIA A100-PCIE-40GB	Off	00000000:3E:00.0	Off			0	
N/A	36C	P0	68W / 250W	20243MiB / 40960MiB	100%	Default	Disabled	
2	NVIDIA A100-PCIE-40GB	Off	00000000:40:00.0	Off			0	
N/A	38C	P0	97W / 250W	20293MiB / 40960MiB	39%	Default	Disabled	
3	NVIDIA A100-PCIE-40GB	Off	00000000:41:00.0	Off			0	
N/A	37C	P0	69W / 250W	20299MiB / 40960MiB	100%	Default	Disabled	
4	NVIDIA A100-PCIE-40GB	Off	00000000:B1:00.0	Off			0	
N/A	35C	P0	63W / 250W	20243MiB / 40960MiB	100%	Default	Disabled	
5	NVIDIA A100-PCIE-40GB	Off	00000000:B2:00.0	Off			0	
N/A	40C	P0	118W / 250W	20293MiB / 40960MiB	38%	Default	Disabled	
6	NVIDIA A100-PCIE-40GB	Off	00000000:B4:00.0	Off			0	
N/A	35C	P0	63W / 250W	20299MiB / 40960MiB	100%	Default	Disabled	
7	NVIDIA A100-PCIE-40GB	Off	00000000:B5:00.0	Off			0	
N/A	36C	P0	64W / 250W	20243MiB / 40960MiB	100%	Default	Disabled	
<hr/>								
<b>Processes:</b>								
GPU	GI	CI	PID	Type	Process name			GPU Memory Usage
ID		ID						
0	N/A	N/A	2784969	C	...anaconda/envs/tg10/bin/python			20418MiB
0	N/A	N/A	2784970	C	...anaconda/envs/tg10/bin/python			618MiB
0	N/A	N/A	2784971	C	...anaconda/envs/tg10/bin/python			618MiB
0	N/A	N/A	2784972	C	...anaconda/envs/tg10/bin/python			618MiB
0	N/A	N/A	2784973	C	...anaconda/envs/tg10/bin/python			618MiB
0	N/A	N/A	2784974	C	...anaconda/envs/tg10/bin/python			618MiB
0	N/A	N/A	2784975	C	...anaconda/envs/tg10/bin/python			618MiB
0	N/A	N/A	2784976	C	...anaconda/envs/tg10/bin/python			618MiB
1	N/A	N/A	2784970	C	...anaconda/envs/tg10/bin/python			20234MiB
2	N/A	N/A	2784971	C	...anaconda/envs/tg10/bin/python			20284MiB
3	N/A	N/A	2784972	C	...anaconda/envs/tg10/bin/python			20290MiB
4	N/A	N/A	2784973	C	...anaconda/envs/tg10/bin/python			20234MiB
5	N/A	N/A	2784974	C	...anaconda/envs/tg10/bin/python			20284MiB
6	N/A	N/A	2784975	C	...anaconda/envs/tg10/bin/python			20290MiB
7	N/A	N/A	2784976	C	...anaconda/envs/tg10/bin/python			20234MiB

## 4bit评估结果

The terminal window displays the following directory structure and the contents of the `all_results.json` file:

```
./TANGOU [SSH: 36.212.4.98]
  .modelscope
  .nv
  .ollama
  .rest-client
  .swanlab
  .trae-server
  .triton
  .vscode-server
  tangou1
  tangou2
    cache
    data
    LLaMA-Factory
    results
      export
      qwen2.5_vl_lora_sft_dev
        test/qwen2.5_vl_lora_sft_dev_eval_test1
          all_results.json (highlighted)
          generated_predictions.json
          predict_results.json
        results_new
        swanlog
        build01.log
        EETQ-1.0.1-cp310-cp310-linux_x86_64.whl
        qwen2.5vl_lora_sft_3_evaluation_no_lora.yaml
        qwen2.5vl_lora_sft_3_evaluation.yaml
```

The `all_results.json` file contains the following evaluation metrics:

```
1  [
2    "predict_bleu-4": 3.8703625000000006,
3    "predict_model_preparation_time": 0.0129,
4    "predict_rouge-1": 15.129037499999999,
5    "predict_rouge-2": 1.9847875,
6    "predict_rouge-l": 6.5945,
7    "predict_runtime": 87.1902,
8    "predict_samples_per_second": 0.069,
9    "predict_steps_per_second": 0.011
10 ]
```

测评， 32B-4bit量化， 7B全量。

数据集coco\_2014\_caption, 部分

60, 30, 10

10 epoch

The screenshot shows a terminal window with two panes. The left pane is a file explorer titled '资源管理器' showing the directory structure of a project named 'TANGOU'. The right pane is a code editor showing the contents of 'dataset\_info.json'.

**Resource Manager (Left):**

- ✓ TANGOU [SSH: 36.212.4.98]
  - ✓ tangou2
    - > cache
    - > configs
    - ✓ data
      - > coco\_2014\_caption
      - > mllm\_demo\_data
        - { } data\_part1\_60.json
        - { } data\_part2\_30.json
        - { } data\_part3\_10.json
        - { } dataset\_info.json
    - { } mllm\_demo.json
  - ✓ LLaMA-Factory
    - > .github
    - > assets
    - > cache
    - > data
    - > docker
    - > evaluation
    - > examples
    - > saves
    - > scripts
    - > src
    - > swanlog
    - > tests
    - ✓ .dockerignore
    - ✓ .env.local
    - ✓ .gitattributes
    - ✓ .gitignore
    - ✓ .pre-commit-config.yaml
    - ✓ CITATION.cff

7B效果

```
conda activate tg10
# 7B, 微调, per_device_train_batch_size: 8, 8卡每张显存占用 28-35G, 1 epoch/30s
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_7B_train.yaml
# 7B微调前, 评估, per_device_train_batch_size: 4, 8卡每张显存占用 26G
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_7B_evaluation_before.yaml
# 7B微调后, 评估, per_device_train_batch_size: 4, 8卡每张显存占用 26G
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_7B_evaluation_after.yaml
```

7B微调loss (收敛不明显)

```

swanlab: Tracking run with swanlab version 0.4.12
swanlab: Run data will be saved locally in /home/tangou/tangou2/swanlog/run-20250421_020121-a3b1799d
swanlab: Experiment name: swan-
swanlab: ★ Run 'swanlab watch /home/tangou/tangou2/swanlog' to view SwanLab Experiment Dashboard locally
{'loss': 0.4919, 'learning_rate': 0.0001, 'epoch': 1.0}
{'loss': 0.4874, 'learning_rate': 9.698463103929542e-05, 'epoch': 2.0}
{'loss': 0.4866, 'learning_rate': 8.8302221559489e-05, 'epoch': 3.0}
{'loss': 0.4913, 'learning_rate': 7.500000000000001e-05, 'epoch': 4.0}
{'loss': 0.4887, 'learning_rate': 5.868240888334653e-05, 'epoch': 5.0}
{'loss': 0.4921, 'learning_rate': 4.131759111665349e-05, 'epoch': 6.0}
{'loss': 0.493, 'learning_rate': 2.500000000000001e-05, 'epoch': 7.0}
70%| 12] [WARNING] [stage3.py:2139:step] 1 pytorch allocator cache flushes since last step. this happens when there is high memory pressure and is detrimental to performance. if this is happening frequently consider adjusting settings to reduce memory consumption. If you are unable to make the cache flushes go away consider adding get_accelerator().empty_cache() calls in your training loop to ensure that all ranks flush their caches at the same time
{'loss': 0.4923, 'grad_norm': 3.0219369532993468, 'learning_rate': 1.1697777844051105e-05, 'epoch': 8.0}
{'loss': 0.4861, 'grad_norm': 3.0219369532993468, 'learning_rate': 3.0153689607045845e-06, 'epoch': 9.0}
{'loss': 0.4835, 'grad_norm': 3.0219369532993468, 'learning_rate': 0.0, 'epoch': 10.0}
100%| 10/10 [02:37<00:00, 14.66s/it] [INFO|trainer.py:3942]
2025-04-21 02:04:22,460 => Saving model checkpoint to results/lora/qwen2.5_vl_lora_sft_7B/checkpoint-10
[INFO]configuration_utils.py:697] 2025-04-21 02:04:22,474 => loading configuration file /data/huggingface/hub/models--Qwen--Qwen2.5-VL-7B-Instruct/snapshots/68156fd997cdc9f710620466735af49862bb81f6/config.json
[INFO]configuration_utils.py:711] 2025-04-21 02:04:22,476 => Model config Qwen2.5_VLConfig /

```

## 结果

```

// 7B微调前
{
    "predict_bleu-4": 8.043468749999999,
    "predict_model_preparation_time": 0.0075,
    "predict_rouge-1": 18.08796875,
    "predict_rouge-2": 4.3805499999999995,
    "predict_rouge-l": 11.1585375,
    "predict_runtime": 8.4024,
    "predict_samples_per_second": 3.57,
    "predict_steps_per_second": 0.119
}
// 7B微调后
{
    "predict_bleu-4": 8.093843750000001,
    "predict_model_preparation_time": 0.0101,
    "predict_rouge-1": 18.75373125,
    "predict_rouge-2": 4.851125,
    "predict_rouge-l": 11.525771875,
    "predict_runtime": 12.6196,
    "predict_samples_per_second": 2.377,
    "predict_steps_per_second": 0.079
}

```

## 32B效果

```

conda activate tg10
# 32B, 微调, per_device_train_batch_size: 6, 8卡每张显存占用 40G, 1 epoch/120s
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_32B_train.yaml

# 32B微调前, 评估, 8bit量化, per_device_train_batch_size: 2, 8卡每张显存占用 38G
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_32B_evaluation_before_8bit.yaml
# 32B微调前, 评估, 4bit量化, per_device_train_batch_size: 2, 8卡每张显存占用 24G
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_32B_evaluation_before_4bit.yaml

```

```

# 32B微调后, 评估, 8bit量化, 爆显存
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_32B_evaluation_after_8bit.yaml
# 32B微调后, 评估, 4bit量化, per_device_train_batch_size: 2, 8卡每张显存占用 24G
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_32B_evaluation_after_4bit.yaml

```

## 32B微调loss (收敛不明显)

```

swanlab: Tracking run with swanlab version 0.4.12
swanlab: Run data will be saved locally in /home/tangou/tangou2/swanlog/run-20250421_022749-a3b1799d
swanlab: Experiment_name: swan-
swanlab: ★ Run `swanlab watch /home/tangou/tangou2/swanlog` to view SwanLab Experiment Dashboard locally
{'loss': 1.059, 'learning_rate': 0.0001, 'epoch': 1.0}
{'loss': 1.052, 'learning_rate': 9.698463103929542e-05, 'epoch': 2.0}
{'loss': 1.0609, 'learning_rate': 8.83022221559489e-05, 'epoch': 3.0}
[30%]
[2025-04-21 02:34:55,887] [WARNING] [stage3.py:2139:step] 1 pytorch allocator cache flushes since last step. th
s happening frequently consider adjusting settings to reduce memory consumption. If you are unable to make the
ng loop to ensure that all ranks flush their caches at the same time
{'loss': 1.0639, 'grad_norm': 1.4616900753105404, 'learning_rate': 7.500000000000001e-05, 'epoch': 4.0}
{'loss': 1.0489, 'grad_norm': 1.4616900753105404, 'learning_rate': 5.868240888334653e-05, 'epoch': 5.0}
{'loss': 1.0447, 'grad_norm': 1.4616900753105404, 'learning_rate': 4.131759111665349e-05, 'epoch': 6.0}
{'loss': 1.0474, 'grad_norm': 1.4616900753105404, 'learning_rate': 2.500000000000001e-05, 'epoch': 7.0}
{'loss': 1.0382, 'grad_norm': 1.7393602535478587, 'learning_rate': 1.169777844051105e-05, 'epoch': 8.0}
{'loss': 1.0258, 'grad_norm': 1.7393602535478587, 'learning_rate': 3.0153689607045845e-06, 'epoch': 9.0}
{'loss': 1.0434, 'grad_norm': 1.7393602535478587, 'learning_rate': 0.0, 'epoch': 10.0}
100%

```

## 结果

```

// 32B微调前, 8bit
{
    "predict_bleu-4": 4.050425000000001,
    "predict_model_preparation_time": 0.0026,
    "predict_rouge-1": 11.830003125,
    "predict_rouge-2": 2.671021875,
    "predict_rouge-l": 5.869315625,
    "predict_runtime": 236.453,
    "predict_samples_per_second": 0.127,
    "predict_steps_per_second": 0.008
}

// 32B微调前, 4bit
{
    "predict_bleu-4": 3.984387499999995,
    "predict_model_preparation_time": 0.0026,
    "predict_rouge-1": 11.7092875,
    "predict_rouge-2": 2.0235125,
    "predict_rouge-l": 5.531334375,
    "predict_runtime": 193.3049,
    "predict_samples_per_second": 0.155,
    "predict_steps_per_second": 0.01
}

// 32B微调后, 4bit
{
    "predict_bleu-4": 4.27783125,
    "predict_model_preparation_time": 0.0129,

```

```
"predict_rouge-1": 12.3642375,  
"predict_rouge-2": 2.23608125,  
"predict_rouge-l": 6.093603125,  
"predict_runtime": 220.0491,  
"predict_samples_per_second": 0.136,  
"predict_steps_per_second": 0.009  
}
```