

# 本 科 期 末 论 文

# （主 修 专 业）

山河绘卷

—人工智能期末项目

姓 名：杨雨铮 22920212204479

郑凯翔 34520212201795

赵原 30320202203216

庄永欣 22920212204512

徐晨 11920202204430

学 院：信息学院

专 业：数字媒体技术年 级：2021 级

校内指导教师：姚俊峰教授

二〇二四年六月 二十 日

# Abstract

***The project combines static Chinese landscape paintings with AI technology to form Chinese landscape animations. Through intelligent recognition and analysis of flowers, trees, clouds, water, etc. in landscape paintings, streams can flow and flowers and birds can dance, bringing more visual impact to users and allowing them to experience the unique charm and appeal of Chinese painting art more immersively .***

***In the system, users input a static picture, and through the visual interactive interface, the system can appreciate the style of the landscape painting. Users can understand the style and artistic considerations of the landscape painting. At the same time, they can also write a story. According to the content of the painting, a relevant short story can be written to help users understand the theme and connotation of the painting. In addition, the painting can be scored, and the aesthetic evaluation function can be used to analyze whether the painting conforms to human aesthetics.***

***On the technical level, this project mainly uses MiniGPT-4, which aims to align the visual information from the pre-trained visual encoder with the advanced large language model (LLM). It is trained to achieve image description for Chinese style landscape paintings, and then Qwen is used as a natural language model to analyze and process the results obtained by the multi-modal model to obtain the subsequent promotion . The picture and prompt are handed over to the final Image-to-reward cross-modal model to obtain the final video result. Multiple large models are interspersed to achieve functions such as image aesthetic evaluation and image quality detection.***

***Finally, we integrated the large models used in the entire system and built a complete system with front-end and back-end. Users only need to input pictures in the specified area to get all the necessary outputs. In short, we created an open system that can help the public appreciate aesthetic pictures with a low threshold.***

**key words:** Transformerr MLLM Image description Aesthetic evaluation Image generation video Chinese landscape paintings

# Content:

[Abstract 2](#_bookmark0)

1. [Introduction 5](#_bookmark1)
   1. [Background 5](#_bookmark2)
      1. [Chinese painting and landscape painting 5](#_bookmark3)
      2. [The inheritance of Chinese painting 6](#_bookmark4)
      3. [Chinese painting and the general public 10](#_bookmark5)
   2. [Integration of traditional Chinese painting and new technologies 11](#_bookmark6)
   3. [Project Purpose 14](#_bookmark7)
   4. [The significance of the projects 16](#_bookmark8)
   5. [Project Introduction 17](#_bookmark9)
2. [Related Work 18](#_bookmark10)
   1. [Large Model Survey Comparison 18](#_bookmark11)
      1. [Image Description 19](#_bookmark12)
      2. [Image To video 21](#_bookmark13)
   2. [Dataset Collection and Cleaning 28](#_bookmark14)
      1. [Overview 28](#_bookmark15)
      2. [Dataset Collection 29](#_bookmark16)
      3. [Dataset processing and cleaning 31](#_bookmark17)
   3. [Model deployment 36](#_bookmark18)
   4. [Front-end and back-end configuration 36](#_bookmark19)
      1. [Gradio 36](#_bookmark20)
      2. [PyQt 37](#_bookmark21)
3. [Method 39](#_bookmark22)
   1. [MiniGPT-4 39](#_bookmark23)
      1. [detail 39](#_bookmark24)
      2. [method 40](#_bookmark25)
      3. [LLM and Vicuna 41](#_bookmark26)
      4. [VIT Backbone Structure 42](#_bookmark27)
   2. [modelscope 44](#_bookmark28)
      1. [Qwen 44](#_bookmark29)
      2. [Image quality damage analysis 47](#_bookmark30)
   3. [Image Aesthetic Assessment 47](#_bookmark31)
      1. [Image-reward 47](#_bookmark32)
      2. [CLIP+MLP Aesthetic Score Predictor 48](#_bookmark33)
   4. [Image-to-video 49](#_bookmark34)
      1. [ST -UNet 49](#_bookmark35)
      2. [VLDM 53](#_bookmark36)
   5. [Fine-tuning and Dataset Cleansing Approach from InstructionGPT-4 for Mini-GPT4[4] . 57](#_bookmark37)
4. [Experiment contents 64](#_bookmark38)
   1. [Dataset Processing Algorithm 64](#_bookmark39)
      1. [DataSelector.py algorithm idea 64](#_bookmark40)
      2. [StyleClassification.py algorithm ideas 66](#_bookmark41)
      3. [Watermark Removal 73](#_bookmark42)
      4. [Image-Video Collection Cleaning 74](#_bookmark43)
   2. [Front-end and back-end connectivity calling algorithm 76](#_bookmark44)
   3. [Back-end 80](#_bookmark45)
      1. [Functionality 80](#_bookmark46)
      2. [Detailed Design 81](#_bookmark47)
   4. [Model Function and Training 82](#_bookmark48)
5. [Conclusion 83](#_bookmark49)
6. [References 83](#_bookmark50)

## Introduction

### Background

#### Chinese painting and landscape painting

Chinese painting is a traditional Chinese painting form, which is to use a brush dipped in water, ink, and color to paint on silk or paper. Tools and materials include brushes, ink, Chinese painting pigments, rice paper, silk, etc. The subject matter can be divided into figures, landscapes, flowers and birds, etc., and the techniques can be divided into figurative and freehand. In terms of content and artistic creation, Chinese painting reflects the ancient people's understanding of nature, society, and related politics, philosophy, religion, morality, literature and art.

Chinese landscape painting is the most profound sedimentation of Chinese people's emotions. The mainland cultural consciousness of traveling and enjoying the mountains and rivers, the inner cultivation consciousness of taking mountains as virtue and water as nature, and the visual illusion consciousness of being close yet far away have always been the central axis of landscape painting. From landscape painting, we can appreciate the artistic conception, style, charm and color of Chinese painting. No other painting can give Chinese people more emotions like landscape painting . If we talk about scriptures and debates with others, landscape painting is the national heritage, the confidence of classics, the image of the country, and the temperament of people.

In terms of ideological content and artistic creation, Chinese painting reflects the social consciousness and aesthetic taste of the Chinese nation, and embodies the Chinese people's understanding of nature, society, and the politics, philosophy, religion, morality, literature and art related to it. Chinese painting emphasizes

"learning from nature outside and getting the source of the heart inside", and requires "the idea is in the brush first, and the painting is full of the idea". It emphasizes the integration of the self and the object, the creation of artistic conception, and the realization of the spirit with the form, the form and the spirit, and the vivid charm. Because calligraphy and painting have the same origin, and both are closely connected with the bone method of brushwork and the operation of lines in expressing ideas and emotions, painting, calligraphy and seal carving have influenced each other and formed significant artistic characteristics. The tools and materials for painting are specially made pens, ink, paper and silk. Modern Chinese painting has made breakthroughs and developments in inheriting traditions and absorbing foreign techniques.



#### The inheritance of Chinese painting

20th century, and the Chinese painting world can be described as "letting a hundred flowers bloom and a hundred schools of thought contend". Chinese painting is even more colorful in its expression, showing a diversified and comprehensive trend. In the face of complex aesthetic needs, it has become a major issue facing the study of Chinese painting today to understand the current situation and problems of

the Chinese painting world, to clearly recognize the status of Chinese painting art, to understand the core value of Chinese painting, and how to learn Chinese painting in modern society.

The master-apprentice inheritance of Chinese painting has a long history, which can be traced back to ancient times, namely the master-apprentice inheritance system. There are generally two types of master-apprentice relationships: one is the imparting of knowledge to relatives with blood ties; the other is a non-relative relationship without blood ties, but after a long period of close companionship and learning, a sectarian master-apprentice relationship is formed.

Chinese painting seems to be stuck in the quagmire of technology. In the face of the artistic peaks that the ancients have already reached, modern people often just imitate, rather than innovate. This lack of innovative spirit has caused Chinese painting to gradually lose its voice in the torrent of the times. At the same time, with the development of photography technology, Western art forms such as oil painting are also facing similar challenges. However, through the innovation of schools such as Impressionism and Fauvism, the Western art world has successfully freed painting from the shadow of photography and formed its own unique artistic language.



The development of contemporary art is also eroding the living space of traditional Chinese painting. Under the impact of new art forms such as installation art and performance art, many artists began to blindly imitate foreign countries and pursue novel and avant-garde art forms. This kind of plagiarism and imitation not only loses the connotation and value of art itself, but also seriously hinders the inheritance and development of traditional art forms such as traditional Chinese painting.



The inheritance of Chinese painting is a dynamic process of creative transformation, which must focus on the characteristics of the times and the requirements of development. Therefore, the creation of Chinese painting in the new era should not only select the cultural essence that adapts to the times and give it the connotation of the new era, but also supplement, enhance, and improve the connotation and extension of Chinese painting to enhance its influence. Only by combining the times and taking into account the creative transformation of the value concept of reality can the innovation of Chinese painting be achieved, and then the quality of Chinese painting can be improved. Of course, the understanding of tradition and innovation in Chinese painting cannot be solved by reading more books or reciting classic painting theories. The key lies in grasping the essence.

The development of Chinese painting is a dynamic system, and its inheritance and innovation complement each other. We must innovate in inheritance and reform and enrich tradition in practice. Therefore, Chinese painting is related to both the present and the future. Chinese painters of all dynasties have shown the aesthetic

consciousness of different cultures according to the political and cultural environment and style of the time, giving vitality to paintings. It can be seen that Chinese painting carries the sedimentation of national consciousness for thousands of years, and objectively shows the development process of Chinese painting. Although the social background and living environment are constantly changing, the cultural values of Chinese painting are still maintained, and a large number of excellent works have been precipitated. In this way, Chinese painters should not only deepen their understanding of tradition, but also effectively transform it into something new. Only in this way can we ensure that the roots are continuous, face the world, and demonstrate the creativity of Chinese painting that inherits the past and opens up the present.

#### Chinese painting and the general public

Traditional Chinese paintings often use symbolic imagery and metaphors, which require viewers to have a certain understanding and knowledge of Chinese culture, history, and traditional symbols, which may be confusing for viewers who are not familiar with Chinese culture.

Secondly, traditional Chinese paintings focus on the expression of artistic conception and atmosphere rather than direct visual presentation. The audience needs to feel the emotions and themes conveyed by the paintings through the delicate brushwork, composition, and use of lines, which requires a certain aesthetic sensitivity and artistic appreciation.

In addition, traditional Chinese paintings often use simple, blank expression techniques, using images and symbols to convey information rather than detailed

descriptions. This expression method requires the audience to actively associate and think to discover the deep meaning of the painting, which may require a certain amount of time and patience for audiences who are accustomed to intuitive and clear information.

However, despite the difficulty in appreciating traditional Chinese paintings, they have unique charm and value. Through in-depth study and understanding of Chinese culture, and through dialogue and interpretation with artists, the audience can gradually comprehend and appreciate the spiritual connotation and beauty contained in traditional Chinese paintings. In this process, the audience will also experience the pleasure and enlightenment of dialogue with traditional culture, and expand their

artistic vision and aesthetic realm.



### Integration of traditional Chinese painting and new technologies

As a traditional art form with a history of thousands of years, Chinese traditional

painting carries rich cultural connotations and aesthetic values. However, with the rapid development of science and technology, traditional art is facing new challenges and opportunities. The development of science and technology has had a profound impact on the inheritance of Chinese painting, injecting new vitality and innovation into traditional art, and bringing about a series of thinking and exploration.

First, the advancement of science and technology has provided a way for the digital preservation and dissemination of traditional Chinese painting. Traditional Chinese paintings are precious and fragile, and their long-term preservation and protection face difficulties. However, high-resolution digital scanning technology allows Chinese paintings to be accurately preserved in digital form, which can not only effectively prevent physical damage, but also be widely disseminated and shared on the Internet and digital platforms. This provides a broader space and opportunity for the inheritance and promotion of Chinese paintings, allowing more people to appreciate and understand the charm of traditional art.

Secondly, the development of science and technology has provided new tools and media for artistic creation. Digital tools such as digital painting software and drawing boards allow artists to create paintings on computers and modify and edit them at any time. This innovative technology provides artists with greater creative space and inspiration, allowing them to express their artistic concepts and creativity more freely. At the same time, the use of digital technology has also promoted the integration of traditional techniques and modern innovations, injecting new artistic elements and forms of expression into Chinese painting.

Third, the development of science and technology has changed the way of art education, bringing new teaching models and opportunities for the inheritance of Chinese painting. Through distance education platforms and online courses, students

can receive Chinese painting education from all over the world anytime and anywhere. Video teaching and interactive communication allow students to learn and guide with Chinese painting masters in real time, transcending geographical and time constraints. This teaching model not only makes the inheritance of Chinese painting more convenient and extensive, but also provides students with more learning resources and opportunities, inspiring their interest and love for traditional art.

In addition, the application of virtual reality technology also provides the audience with an immersive experience of appreciating traditional Chinese paintings. Through virtual exhibitions and digital art galleries, the audience can appreciate the details and artistic expressions of traditional Chinese paintings up close through head-mounted display devices or smart phones, while obtaining more background information and artistic interpretations. This interactive and immersive experience enables the audience to more deeply understand and feel the cultural connotation and artistic charm of traditional Chinese paintings.

However, although the development of science and technology has brought many opportunities and conveniences to the inheritance of traditional Chinese painting, we must also remain cautious. As a unique art form, the essence and characteristics of traditional Chinese painting should not be weakened by science and technology. In the process of applying science and technology, we need to balance the relationship between traditional techniques and innovative technologies to ensure that the application of science and technology will not destroy the unique charm and cultural connotation of traditional Chinese painting works. Protecting and inheriting traditional manual skills and artistic spirit is still an important task of the inheritance of traditional Chinese painting, and science and technology are only tools to assist and promote it.

Therefore, the relationship between scientific and technological development and the inheritance of traditional Chinese painting is not an opposition, but a relationship of mutual integration and promotion. The advancement of science and technology has injected new vitality and innovation into the inheritance of traditional Chinese painting, but it also requires us to maintain awe and respect for traditional art. Only in the combination of tradition and modernity can traditional Chinese painting shine more brilliantly, and continue to be inherited and developed in contemporary society, bringing people the enjoyment of beauty and the inheritance of culture.

### Project Purpose

As an important part of human culture, art has been constantly developing and innovating. With the advancement of science and technology, especially the rapid development of artificial intelligence (AI) technology, traditional art forms have also ushered in new changes and opportunities. This article will explore how to inject innovative artistic expressions into traditional Chinese painting by combining Chinese landscape painting with AI technology, promote the dissemination and exchange of traditional culture, and explore its potential in technology research, application, and cultural heritage protection and development.

Chinese traditional landscape painting is a treasure in Chinese traditional culture, known for its unique aesthetic and artistic expression. However, with the changing times, traditional forms of expression sometimes have difficulty attracting the attention of the younger generation. By introducing AI technology into Chinese traditional landscape painting, we can inject new life into this traditional art form. AI can generate new landscape paintings through data analysis and deep learning, or make static landscape paintings dynamic, presenting a more vivid and modern artistic

experience to the audience. This not only expands the boundaries of Chinese painting art, but also provides new tools and methods for artistic creation.

By presenting traditional Chinese landscape paintings to the audience in a new form and combining it with modern AI technology, traditional Chinese paintings can be promoted to a wider audience. Through digitalization and network dissemination, traditional Chinese landscape paintings can break through geographical and temporal limitations, allowing more people to have the opportunity to contact and understand this art. At the same time, the application of AI technology can make the dissemination of traditional Chinese paintings more vivid and interesting, enhance the audience's interactive experience, and promote the dissemination and exchange of traditional Chinese culture.

In the process of applying AI technology to traditional Chinese landscape painting, a lot of technical research and practice are needed. By investigating different AI generation models, trying and comparing their effects, we can find the most suitable technical solutions for the dynamicization of traditional Chinese landscape painting and the transfer of ink painting . For example, using advanced technologies such as generative adversarial networks (GAN) and convolutional neural networks (CNN), it is possible to achieve intelligent processing and re-creation of traditional landscape paintings. This not only helps to explore the application prospects of AI technology in the field of art, but also provides a new direction for the development of related technologies.

The protection and inheritance of traditional culture is one of the important topics in today's society. As an important part of Chinese culture, the protection and development of Chinese landscape painting is particularly important. By combining traditional Chinese painting with modern AI technology, we can inject new power into

this cultural heritage and achieve a win-win situation of protection and development. For example, AI technology can be used to digitally archive and restore precious ancient paintings to extend their preservation time; at the same time, AI-generated landscape paintings can also serve as a continuation of traditional art and enrich the connotation of cultural heritage.

The combination of AI technology and traditional Chinese landscape painting is an innovative collision of traditional art and modern technology. This innovation not only injects new forms of expression into traditional Chinese painting art and expands the boundaries of art, but also promotes the dissemination and exchange of traditional culture. Through in-depth technical research and application, we can explore more application prospects of AI in the field of art and contribute to the protection and development of cultural heritage. In the future, with the further development of technology, the integration of traditional art and modern technology will bring more surprises and possibilities.

### The significance of the projects

In today's fast-paced society, there is a growing need for systems that can help people appreciate traditional Chinese landscape paintings in a meaningful way. Our project leverages cutting-edge deep learning technology to provide an innovative solution for this need. By integrating AI, we offer users a unique opportunity to engage with traditional Chinese paintings through dynamic and interactive experiences that highlight the content, significance, and aesthetic value of these artworks.

The system simplifies the appreciation process: users simply upload a painting, and the system guides them through a multi-step analysis. Initially, a picture

description model interprets and analyzes the painting, offering a preliminary appreciation. This analysis is then refined by a natural language model, which provides a deeper, more nuanced interpretation. The system also includes aesthetic evaluation and quality analysis features, offering users a comprehensive understanding of the artwork's beauty and craftsmanship. Ultimately, the parameters collected are used to generate dynamic videos, allowing users to experience the paintings in a lively, animated form.

By transforming static images into dynamic experiences, our system not only makes traditional Chinese landscape paintings more accessible but also enhances users' appreciation and understanding of these cultural treasures.

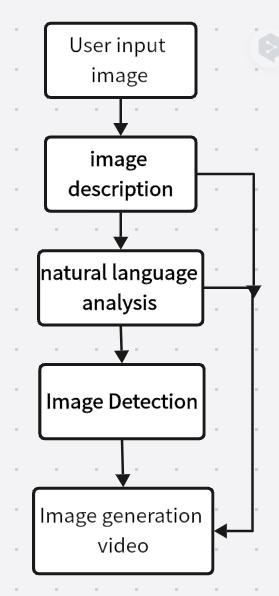
### Project Introduction

As society urgently needs a system to help people appreciate traditional Chinese landscape paintings, we developed this system based on current deep learning technology to help people appreciate paintings from the perspectives of picture content, meaning, and aesthetic value. At the same time, it intuitively presents the beauty of landscape paintings to users in a dynamic form.

In this system, users only need to input the pictures they need to appreciate, and then they can get the output step by step according to the system guidance. This system firstly uses the picture description model to describe and analyze the paintings and make a preliminary appreciation of the paintings. After that, the output of the description model is handed over to the natural language model, and the same processing results are obtained in the form of a fixed prompt, which is fed back to the user as a secondary appreciation result. After that, the pictures are aesthetically evaluated and the quality is analyzed to give the user a direct feeling. Finally, all

parameters are passed to the multi-modal image-generated video model to generate dynamic videos, allowing users to directly feel the beauty of the dynamic paintings.

The specific workflow is as follows:



## Related Work

### Large Model Survey Comparison

At present, there are a large number of multi-modal open source models and large model calling platforms in the industry, with a very rich stock. However, their effects and specific functions are not the same. Therefore, our team first conducted a comparative analysis of some large models investigated, and then comprehensively considered the project development cost, technical feasibility, project quality and other aspects to select the appropriate model.

#### Image Description

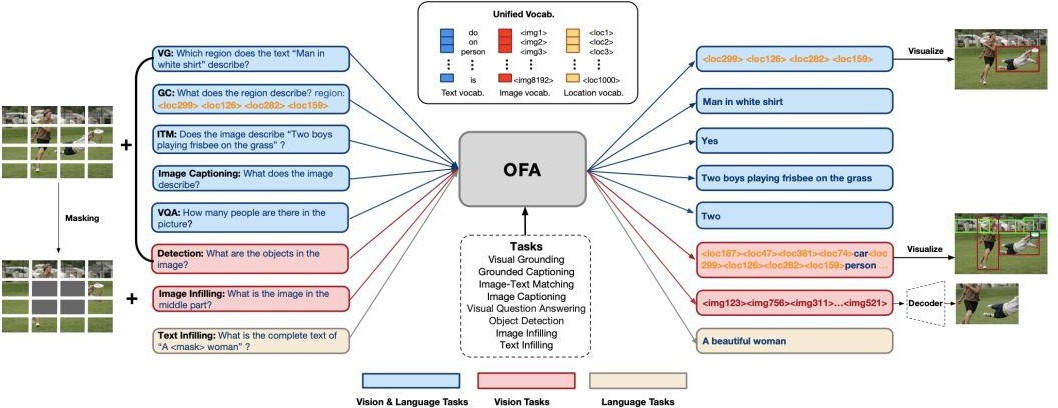
Image description technology is to use images as input, and through mathematical models and calculations, the computer outputs natural language description text corresponding to the image, so that the computer has the ability to "see the picture and speak". It is another new task in the field of image processing after image recognition, image segmentation and target tracking. The development of this field is relatively mature, so there are many models to choose from. Here we mainly choose the OFA multi-modal description model and Mini-GPT4 as comparison objects to show the reasons why our team chose Mini-GPT4 and its advantages.

1. OFA:

It was published by Alibaba DAMO Academy and uses a general multi-modal pre-training model. It uses a simple sequence-to-sequence learning framework to unify the modalities (cross-modal, visual, language, etc.) and tasks (such as image

generation, visual localization, image description, image classification, text generation, etc.) [1] .

The implementation principle of OFA is relatively simple, and the core model architecture is the most classic transformer encoder-decoder. In order to integrate both pre-training and fine-tuning into this architecture, OFA expresses various tasks involving multi-modality and unimodality (i.e., NLP and CV) in the form of Sequence-to-Sequence, and uses the above encoder-decoder model for training. There is no need to add task-specific model layers for pre-training and fine-tuning, such as the classification layer used by BERT in classification task fine-tuning, to reduce the difference between pre-training and fine-tuning. In terms of specific implementation,

OFA has made a series of designs for unified pre-training , including how to implement the input of modal information such as images, texts, detection boxes, etc. with different resolutions, and how to unify different multimodal and single-modal tasks into a sequence-to-sequence format, as shown in the following figure:

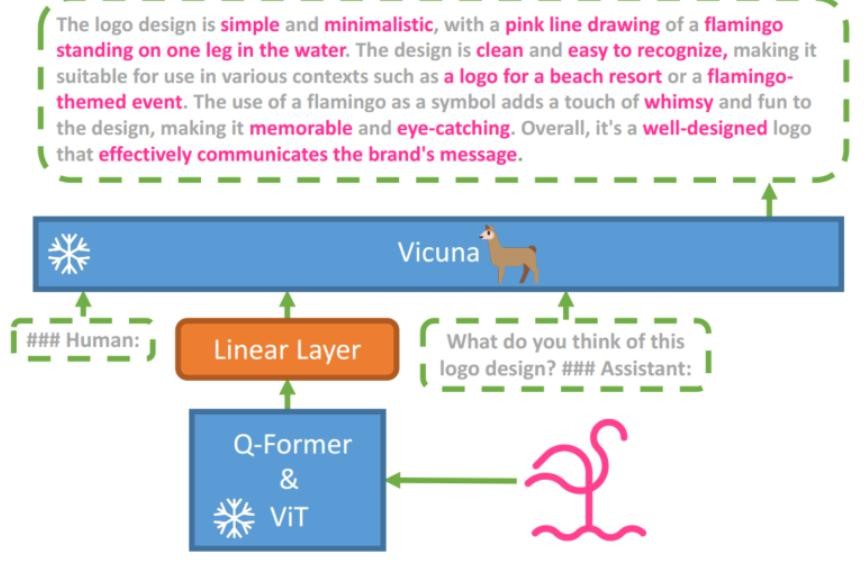
Obviously, OFA is a lightweight multimodal large model that uses a relatively simple model architecture to achieve image description and recognition, and also lays the foundation for subsequent lightweight training and fine-tuning. However, due to the simplicity of the process and the use of only the most classic transformer encoder-decoder, the model itself is not powerful enough to meet the needs of some projects.

1. MiniGPT-4:

GPT-4 exhibits extraordinary multimodal capabilities that are rarely seen in previous work, but the technical details behind GPT-4 remain undisclosed. The authors of MINi GPT-4 believe that the enhanced multimodal generation capabilities stem from the use of complex large language models (LLMs). To study this phenomenon, MiniGPT-4 was proposed, which achieves partial results by aligning a frozen visual encoder with a frozen advanced LLM Vicuna through a mapping layer [ 2 ] .

The main core of minigpt4 is that it uses two stages when training Linear. The first stage uses a low-level data set (5 million pairs) for training, and the second stage manually screens the low-level data set (using GPT to generate text captions for each image - 5,000 pairs, and manually screen out the generated image-text pairs that are more consistent with the image content - 3,500 pairs), and then uses these high-quality data sets to fine-tune the model , thereby achieving better results. This project also uses the same method in fine-tuning training, and fine-tunes the model of

the second stage.



After analyzing the principles of the two and testing them, our team found that MiniGPT-4 is far better than OFA in multimodal image description work. So after weighing the pros and cons, we decided to use MiniGPT-4 for subsequent project development.

#### Image To video

1. Stable Video Diffusion

Stable Video Diffusion is an advanced video synthesis technique that achieves

the conversion from static images to dynamic videos through a latent diffusion model. **[10]**The model is capable of generating high-resolution videos of 14 or 25 frames, and supports functions such as multi-view generation and frame interpolation . The core idea of Stable Video Diffusion is to decompose the video generation task into two stages. First, it uses a diffusion model to gradually transform random noise into an image similar to the input image. This stage is accomplished by gradually adding details, similar to gradually "zooming in" on a picture. Then, in the second stage, the method uses a conditional variational autoencoder ( cVAE ) to transform the generated image sequence into a video. cVAE is a generative model that can learn the characteristics of data distribution and generate new data samples based on specific conditions.

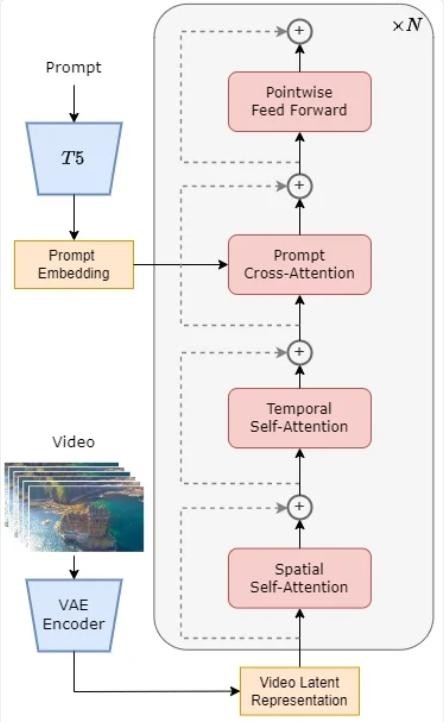


defect:

* + Currently, videos generated directly using SVD are very short, usually within 5 seconds, and cannot yet achieve photo-level realism.
  + The generated results are similar to the early Stable Diffusion, which is relatively uncontrollable and may generate video footage with no motion or very slow motion that deviates greatly from the expected value.
  + It is currently not possible to control the generation of intervention videos through text.
  + Clear text content cannot be presented temporarily.

1. Open-Sora

Open-Sora is a video generation model open sourced by the Colossal-AI team, which aims to replicate OpenAl 's Sora video generation product. Open-Sora is based on the Diffusion Transformer ( DiT ) architecture and is trained in three stages: large-scale image pre-training, large-scale video pre-training , and fine-tuning with high-quality video data to generate video content that matches the text description. The specific design is as follows:**[9]**



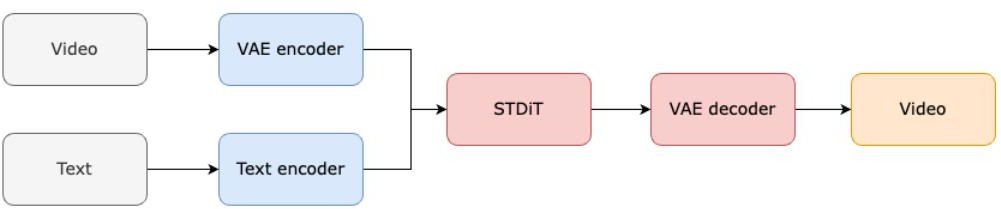
Core Components:

* Pre-trained VAE (Variational Autoencoder): VAE is a component used for data compression, which maps the input video data to a low-dimensional representation in a latent space. In Open-Sora, the encoder part of VAE is used to compress video data during the training phase, while during the inference phase, it samples Gaussian noise from the latent space and generates videos.
* Text Encoder: This component is responsible for converting text cues (such as sentences describing the content of the video) into text embeddings, which are subsequently combined with the video data to ensure that the generated video fits the text description.
* STDiT (Spatial Temporal Diffusion Transformer): This is the core component of Open-Sora, a DiT model that uses the spatial-temporal attention mechanism.

STDiT models the temporal relationship in video data by serially superimposing a one-dimensional temporal attention module on a two-dimensional spatial attention module. In addition, the cross-attention module is used to generate semantic information of the text.

Architecture Design:

* Spatial-temporal attention mechanism: Each layer of the STDiT model contains a spatial attention module and a temporal attention module. The spatial attention module processes the two-dimensional spatial features of the video frame, while the temporal attention module processes the temporal relationship between frames . This design enables the model to effectively handle both spatial and temporal dimensions in video data.
* Cross-Attention: After the temporal attention module, the cross-attention module is used to fuse text embeddings with video features, ensuring that the generated video content matches the text description.
* Training and inference process: In the training phase, the VAE encoder compresses the video data and then uses it together with the text embedding to train the STDiT model. In the inference phase, the noise is sampled from the VAE's latent space and input into the STDiT model together with the text prompt to generate denoised features, which are finally decoded by the VAE decoder to obtain the final video.



1. Image-to-Video HD image generation video large model

The Image-to-Video model used in this project is designed to solve the task of generating high-definition videos based on input images. Image-to-Video is one of the basic models for generating high-definition videos developed by the DAMO Academy. Its core part consists of two stages, which respectively solve the problems of semantic consistency and clarity. The total number of parameters is about 3.7 billion. The model is pre-trained on a large-scale mixture of video and image data and fine-tuned on a small amount of high-quality data. The data is widely distributed and diverse in categories, and the model has good generalization to different data. Compared with existing video generation models, Image-to-Video has obvious advantages in clarity, texture, semantics, and temporal continuity.

Image-to-Video is a video diffusion model (VLDM) based on latent space. It uses a specially designed spatiotemporal UNet (ST- UNet ) to perform spatiotemporal modeling in latent space , and then reconstructs the final video through a decoder. In order to generate 720P video, Image-to-Video is divided into two stages. The first stage is to ensure semantic consistency under low resolution conditions, and the second stage is to use the new VLDM for denoising to improve video resolution and improve temporal and spatial consistency at the same time. Through joint optimization of models, data and training, Image-to-Video has the following main features:

* HD & widescreen, can directly generate 720P (1280\*720) resolution video, and compared with existing open source projects, not only the resolution is effectively improved, but the widescreen video produced can be suitable for more scenes.
* Continuity: Through specific training and reasoning strategies, the stability of video detail generation (in both temporal and spatial dimensions) has been

significantly improved.

* By collecting video data of specific styles for training, the texture of the generated videos is significantly improved. It can generate videos of technology, movie colors, cartoon styles, sketches, etc.
* No watermark. The model is trained with our internal large-scale watermark-free videos/images and fine-tuned on high-quality data. The generated watermark-free videos can be applied to more video platforms and reduce many restrictions.

Many design concepts and details of Image-to-Video (such as the core UNet part) are inherited from the DAMO Academy's publicly available work VideoComposer .

Driven by advances in computation, data scaling, and architectural design, current visual generative models, especially diffusion-based models, have made remarkable progress in automated content creation, enabling designers to generate realistic images or videos from textual prompts as input. These methods typically train a powerful diffusion model conditioned by text on large-scale video-text and image-text datasets,**[11]** achieving unprecedented levels of fidelity and diversity. However, despite impressive progress, the controllability of synthesis systems remains a major challenge, which hinders their practical applications. VideoComposer proposes a new generation paradigm centered on the concept of composability, which is able to synthesize videos under a variety of input conditions, resulting in remarkable flexibility.

For the overall architecture of VideoComposer, firstly, a video is decomposed into three representative factors, namely textual conditions, spatial conditions, and key temporal conditions, and then a latent diffusion model is trained to reconstruct the input video under its conditions. In particular, video-specific motion vectors are introduced as a temporal guide for video synthesis during video synthesis, explicitly

capturing inter-frame dynamics and thus providing direct control over internal motion. To ensure temporal consistency, a unified stc encoder is also proposed, which utilizes a cross-frame attention mechanism to capture the spatiotemporal relations in the sequence input, thereby enhancing the cross-frame consistency of the output video . In addition, the stc encoder serves as an interface that allows efficient and unified utilization of control signals from various conditional sequences.

### Dataset Collection and Cleaning

#### Overview

In order to enhance the performance of the model in the field of traditional Chinese paintings, we need to collect more traditional Chinese paintings data for model training to enhance the adaptability and generalization ability of the model in this field.

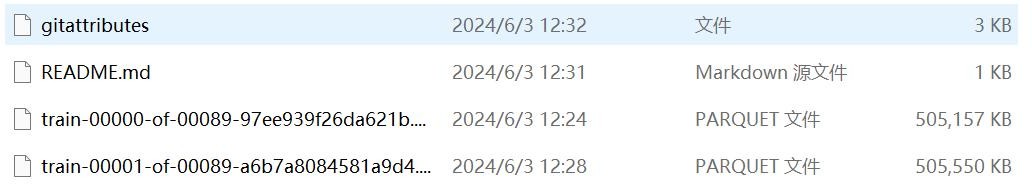
Our dataset collection work mainly focuses on platforms such as Github , Kaggle, and HF Mirror. After collecting the preliminary dataset, we conducted detailed data adaptation and cleaning. In terms of dataset processing , we combined manual annotation and Baidu Smart Cloud AI-assisted annotation, using its advanced image recognition technology to assist in rapid annotation and improve annotation efficiency; in terms of data screening, we combined CLIP Score and Reward Score evaluation indicators to screen the collected paintings. These evaluation indicators can objectively reflect the quality and artistic value of the paintings, helping us to eliminate low-quality works that do not meet the requirements, and ensure that the final dataset is highly representative and accurate.

During the screening process, we paid special attention to retaining data of

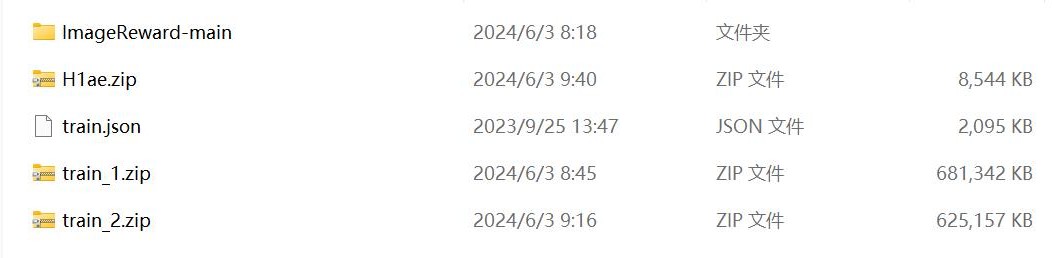
different styles. Therefore, this data cleaning and screening process not only provided us with a high-quality dataset, but also derived the implementation of a simple Chinese style landscape painting style classifier.

#### Dataset Collection

##### Image Description Dataset[5]



1. **Aesthetics Evaluation Dataset[6]**



The dataset for aesthetic evaluation is a subset of the dataset for deploying the model . H1ae is the Chinese landscape painting dataset that we trained and verified.

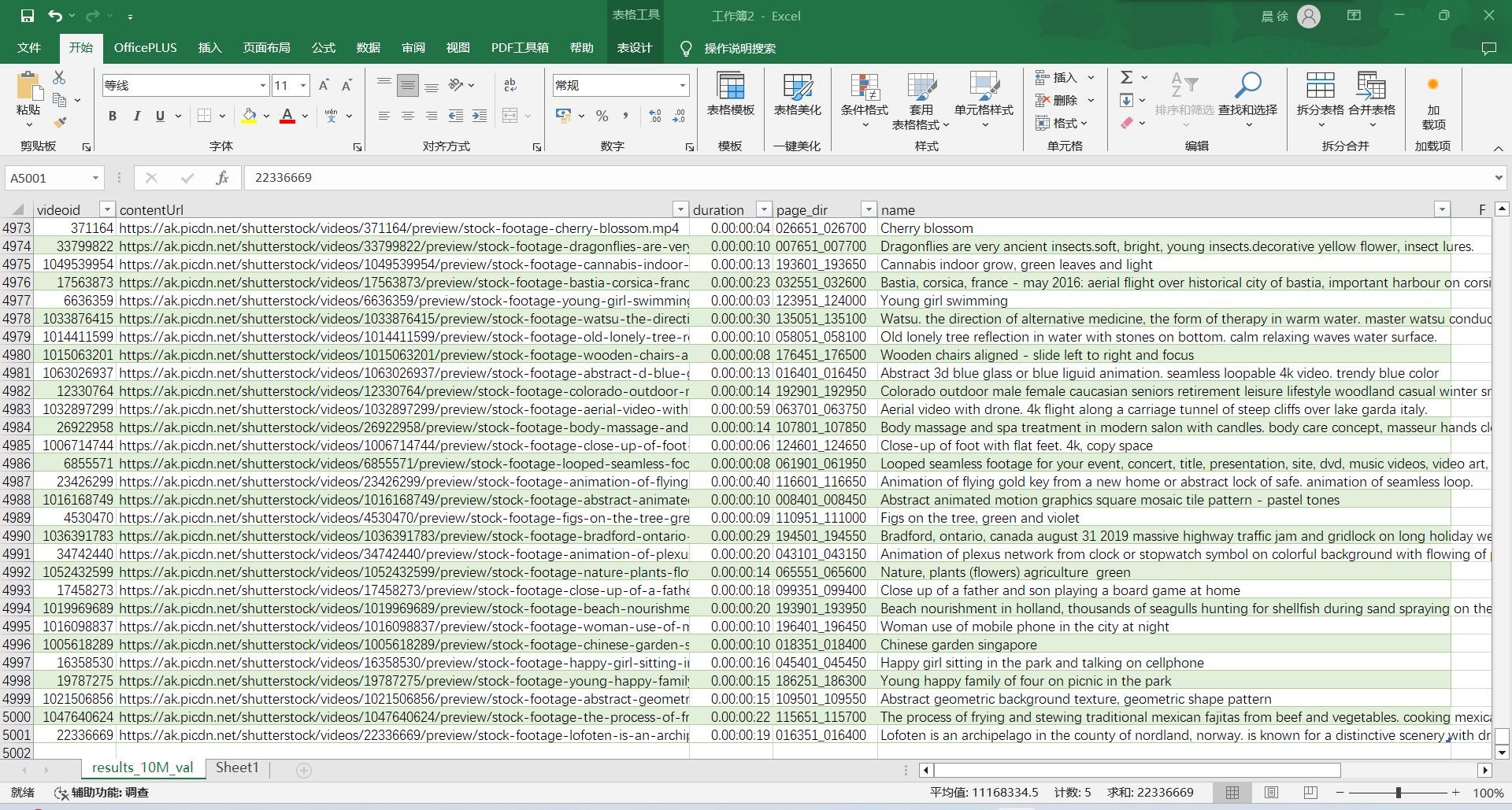
##### Poetry and Ancient Poetry Dataset[8]

A data set of poetry and stories is established, and the model is trained specifically to make the generated poetry closer to the artistic style of the paintings.





##### Image video dataset

WebVid is a large short video-text dataset. (Introduction) WebVid is a text-video dataset scraped from stock footage sites for end-to-end retrieval. WebVid contains 10 million video clips with captions, sourced from the web. The videos are diverse and rich in their content.We selected 5,000 video-text pairs as the video-text dataset.

#### Dataset processing and cleaning

##### Dataset adaptation processing

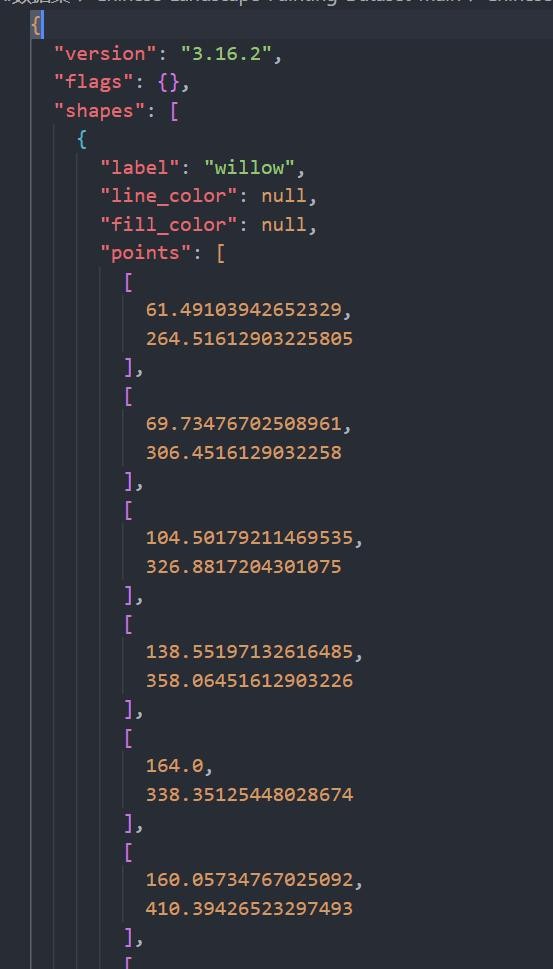
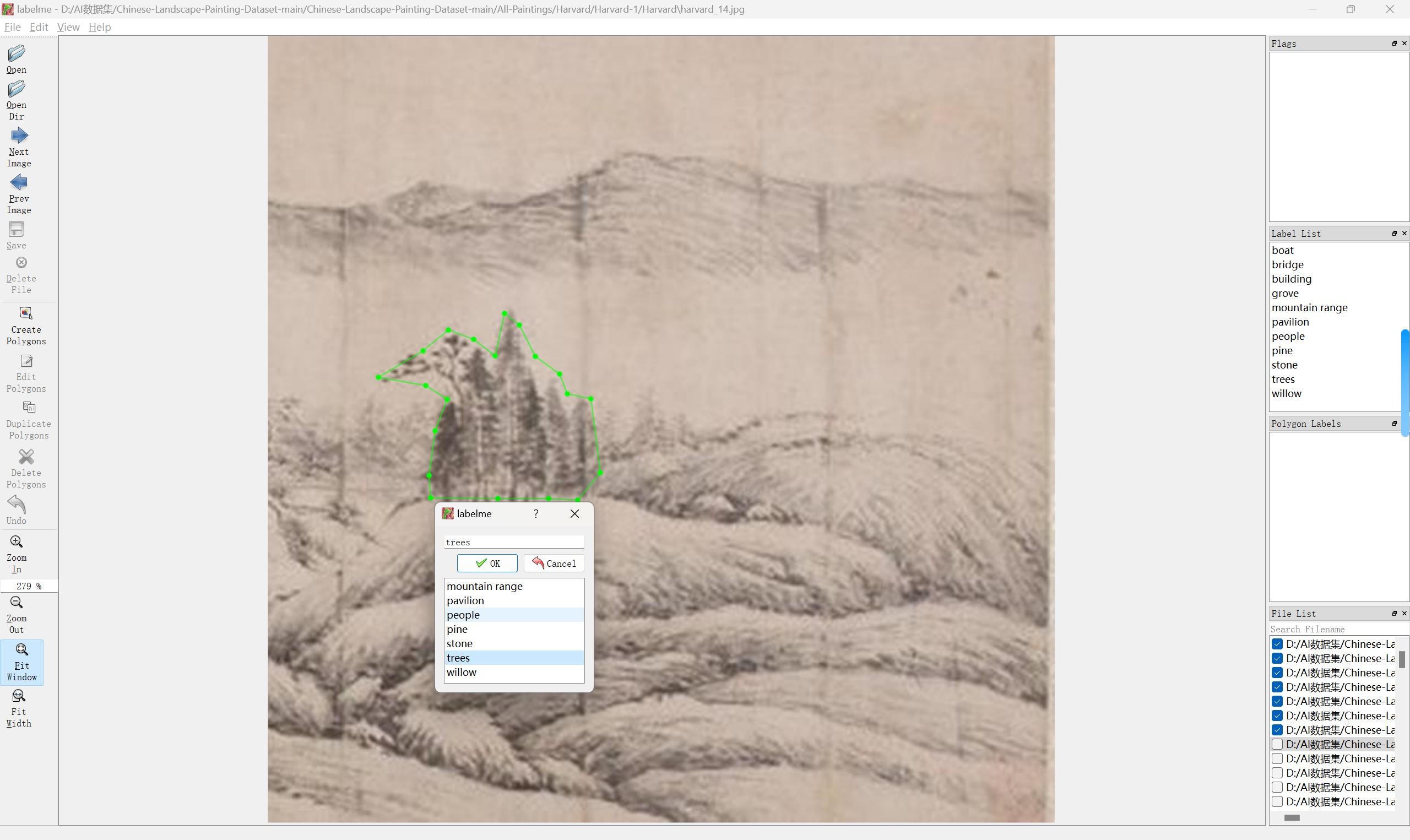
Our model needs to recognize images such as mountains, clouds, boats, trees, waterfalls, birds, etc., and generate descriptive texts in a targeted manner, further animating certain images and outputting video animations. To this end, we optimized and improved the dataset for object recognition and trained the model to make it more accurate in recognizing images in landscape paintings (this part is used for the second stage training of Mini-GPT4):

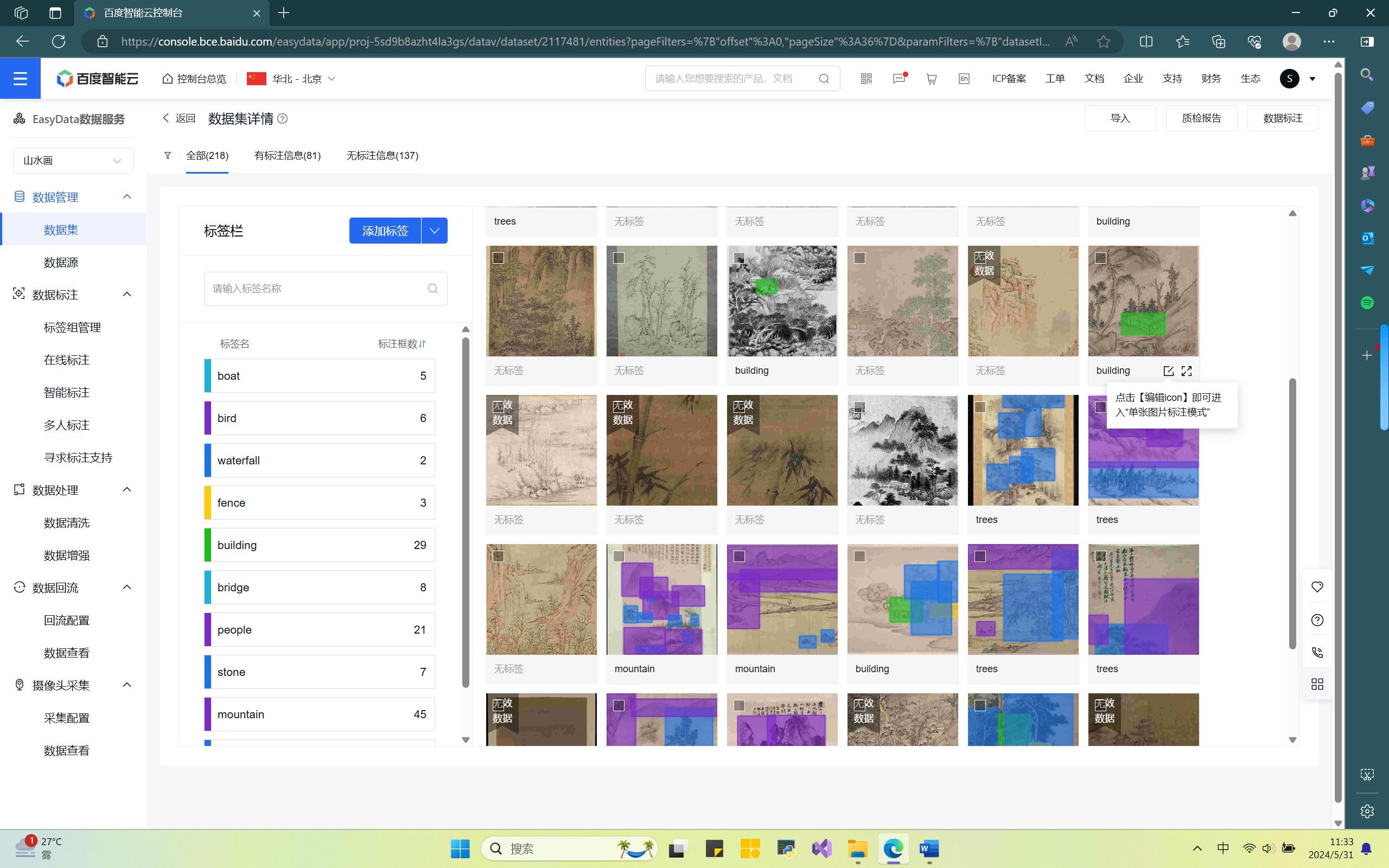
The quality of the dataset was initially ensured by removing some invalid and contaminated data. Finally, a dataset of 1,954 annotated Chinese landscape paintings was obtained,**[7]** which will be used for training and verification of image recognition and appreciation tasks . Considering the small size of the data, we only left 137 unlabeled datasets for testing. The test results were good, but the possibility of overfitting of this type of data (2,000 Chinese landscape paintings in the dataset) cannot be ruled out.

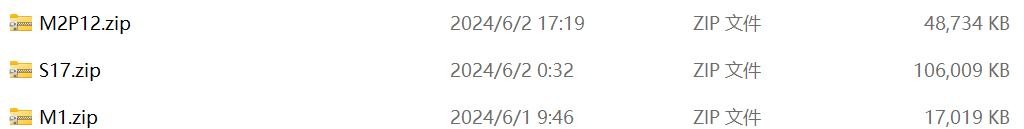
##### Process Description

First, I considered using Labelme to manually label, but found that the workload was too large.

Then we explored Baidu Smart Cloud AI-assisted labeling . After three rounds of difficult example screening and labeling, the data labeling effect was good. The results were uniformly exported in the JSON format of the coco dataset.





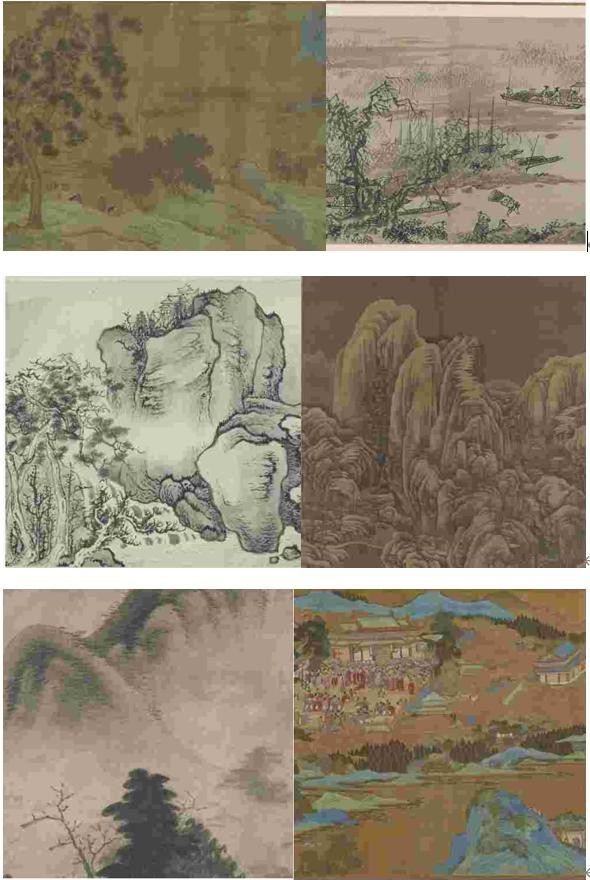


##### Data set cleaning and optimization

After an in-depth study of the paper " InstructionGPT-4: A 200-Instruction

Paradigm for Fine-Tuning MiniGPT-4", we realized that fine-tuning the model using a small amount of high-quality data sets can significantly improve model performance. Therefore, we performed a series of detailed cleaning and optimization work on the second-stage data set.

Firstly, to ensure that the optimized dataset still maintains diversity, we need to weigh the data of different categories. The paper adopts the unsupervised learning method k-balanced classification . Since we have no previous knowledge of style classification and do not know which feature value to use as the classification standard, we use manual labeling to classify the ancient landscape paintings in the dataset. The classification categories are as follows:





During this process, we annotated a total of 1,519 images to ensure that the themes of ancient landscape paintings in the dataset were clear. Subsequently, based on these annotated data, we expanded and trained a style classifier based on color histogram and gray-level co-occurrence matrix. Limited by the non-authoritativeness and scale of the dataset, as well as the limitations of the feature selection of the classifier model (only two eigenvalues were used and there was a lack of deep optimization), our classifier achieved an accuracy of 65.97% in the six-category classification task.

During the data cleaning phase, we selected data with higher CLIP scores and reward scores based on the characteristics of each category. These parameters can comprehensively evaluate the quality and relevance of the data, ensuring that the data we selected not only fits the theme of ancient landscape paintings, but also has high quality. We carefully selected high-quality triplet data from each category and combined them to form a final high-quality dataset.

##### Training of MiniGPT-4

Pre-training datasets by merging visual and language data (i.e., image-text pairs), visual data (i.e., raw image data, object labeling data), and language data (i.e., plain text) . For replication, only publicly available datasets are used. Pre-training data is carefully filtered to exclude images that appear in downstream task validation and test sets to avoid data leakage.

* + The first stage of pre-training:

In the initial pre-training stage, the model aims to capture visual-linguistic knowledge from a large number of aligned image-text pairs. The output injected into the projection layer is regarded as a soft hint to the LLM, prompting it to generate the corresponding ground truth text.

During the entire pre-training process, the pre-trained visual encoder and LLM are in a frozen state, and only the linear projection layer is pre-trained.

After the first pre-training phase, MiniGPT-4 demonstrated the ability to possess rich knowledge and provide reasonable responses to human queries . However, it had difficulty producing coherent language output, such as producing repeated words or sentences, fragmented sentences, or irrelevant content. These problems hindered MiniGPT-4's ability to have fluent visual conversations with humans.

In order to achieve greater naturalness in the generated language and enhance the usability of the model, the second-stage alignment process is essential. Therefore, a high-quality image-text dataset is carefully curated, specifically tailored for alignment purposes. This dataset is then used to fine-tune MiniGPT-4 in the second-stage alignment process.

After completing the post-processing stage, each image description is manually verified for correctness to ensure its high quality.

* + Second stage fine-tuning:

pre-trained model is fine-tuned with carefully curated high-quality image-text .

During fine-tuning, predefined prompts are used in the following template:



In this prompt, we represent instructions randomly sampled from a predefined set of instructions, which include various forms of instructions such as "Describe this image in detail" or "Can you describe the contents of this image for me ?"

Utilizing only the model after the first pre-training stage may lead to failures such as repeated words or sentences, fragmented sentences, or irrelevant content, which can be alleviated through the second stage fine-tuning process.

As a result, MiniGPT-4 is now able to produce more natural and reliable responses.

Our team conducted the second phase of fine-tuning training on MiniGPT-4 using the dataset they collected and processed.

### Model deployment

In the system developed by our team, multiple large models are used in parallel, which places very high demands on server configuration. Therefore, our team chose to deploy the model on a remote server, build an SSH server connection, and use address penetration technology so that the local machine can obtain the same IP address as the remote server, facilitating coordinated calls between the front-end and back-end.

In addition, all the models used in this system are developed based on the Linux system and deployed using the command line. In particular, for the natural language model Qwen, image quality detection, and image-generated video models, our team used modelscope to deploy and call the models. This approach not only reduces memory loss, but also greatly improves the scalability of the system.

Finally, for the aesthetic evaluation model, our team improved and extracted some existing aesthetic evaluation algorithms. It is no longer a simple deployment.

### Front-end and back-end configuration

#### Gradio

Gradio is a Python library for building interactive interfaces. It helps you quickly create customized web interfaces for displaying and testing machine learning models, natural language processing models, computer vision models, etc.

Gradio makes it very easy to build interactive interfaces without writing tedious

HTML, CSS, and JavaScript code. You can use Gradio to create an interface with input fields (such as text input or image upload) and output fields (such as model prediction results), and users can interact directly with your model.

Gradio supports a variety of input and output types, including text, images, audio, and video. You can process input by defining callback functions and return output to the user. Gradio also provides automatic interface layout and styling, making interface design simple and intuitive.

Gradio 's advantage lies in its ease of use. Its code structure is relatively simple. You only need to define the input and output interfaces to quickly build a simple interactive page, making it easier to deploy models. It is suitable for developers who want to quickly deploy applications in relatively simple scenarios. At the same time, it is easy to share. Gradio can set the share=True parameter when starting the application to create an external sharing link (the link can last for 72 hours, and if it is permanent, it can be deployed on hugging Face), which can be directly shared with users in WeChat.

#### PyQt

PyQt is a Python binding library for creating desktop applications that combines the Qt application framework with the Python programming language. Qt is a cross-platform application development framework that provides a rich set of user interface components and tools for creating powerful and attractive graphical user interface (GUI) applications.

Here are some of the main features and components of PyQt :

Cross-platform: PyQt is cross-platform and supports running on multiple operating systems, including Windows, MacOS, Linux, etc.

Rich components: PyQt provides a wealth of GUI components, such as buttons, text boxes, check boxes, list boxes, menus, etc., which can easily create interactive user interfaces.

Event-driven programming: PyQt uses an event-driven programming model, which triggers corresponding functions or methods by responding to user operations and system events.

Layout manager: PyQt provides a variety of layout managers, such as horizontal layout, vertical layout, grid layout, etc., which are used to automatically adjust and arrange GUI components.

Signal and slot mechanism: The core of PyQt is the signal and slot mechanism, which provides a flexible way to handle events and implement communication between components.

Multimedia support: PyQt supports multimedia functions and can play audio and video files in applications.

Database access: PyQt provides access support to SQLite and other databases, allowing developers to easily store and retrieve data.

Visual design tool: PyQt comes with Qt Designer, a visual design tool that allows developers to create GUI interfaces by dragging and dropping and generate corresponding PyQt code.

In summary, PyQt is a powerful and flexible tool that makes it easy and convenient to develop cross-platform graphical user interface applications using Python. Both beginners and experienced developers can create attractive and interactive applications with PyQt .

## Method

### MiniGPT-4

Leveraging pre-trained large language models ( LLMs) for vision-language tasks has become a clear trend in recent years. Various studies have highlighted the advantages of using autoregressive language models as decoders in vision-language tasks. This approach leverages cross-modal transfer and promotes knowledge sharing between language and multimodal domains.

#### detail

* + - * Pioneering works such as VisualGPT and Frozen have demonstrated the advantages of using a pre-trained language model as a vision-language model decoder.
      * Further progress includes the development of Flamingo, which aligns a pre-trained visual encoder and a language model via gated criss-cross attention , demonstrating impressive contextual few-shot learning capabilities.
      * BLIP-2 combines Flan-T5 and Q-Former to effectively align visual features with language models.
      * PaLM -E focuses on integrating real-world sensor modalities into LLM, thereby establishing a link between real-world perception and human language.
      * shows enhanced visual understanding and reasoning capabilities after pre-training on a large amount of aligned image-text data . LLMs such as ChatGPT have been proven to be an important tool for improving the performance of visual-language tasks.
      * Visual ChatGPT and MM-REACT demonstrate the ability of ChatGPT to coordinate with different visual base models, promoting their collaboration to

tackle more complex challenges.

* + - * ChatCaptioner regards ChatGPT as a questioner, providing BLIP-2 with various questions to answer, and effectively summarizing the image content through multiple rounds of dialogue.
      * Video ChatCaptioner extends this approach to video spatiotemporal understanding.
      * ViperGPT demonstrates the potential of combining LLM with different vision models to programmatically solve complex visual queries.
      * MiniGPT-4 directly aligns visual information with a language model to accomplish various vision-language tasks without using an external vision model.

The goal of MiniGPT-4 is to align the visual information of a pre-trained visual encoder with a large language model (LLM).

* + - * Language Decoder : Using Vicuna as a language decoder, it can perform various complex language tasks.
      * Visual Perception : Use the same visual encoder ViT backbone structure as BLIP-2, combined with a pre-trained Q-Former.
      * Gap Bridging : A linear projection layer is used to bridge the gap between the visual encoder and the LLM.
      * Two-stage training : A two-stage training method is adopted. First, pre-training is performed on aligned image-text pairs to obtain visual language knowledge. Then the pre-trained model is fine-tuned on a smaller but high-quality image-text dataset, using designed dialogue templates to enhance generation reliability and usability.

#### method

MiniGPT-4 is to align the visual information of a pre-trained visual encoder with a large language model (LLM) .

* + - * Language decoder: Vicuna is used as a language decoder, which can perform various complex language tasks.
      * Visual perception: Use the same visual encoder ViT backbone structure as BLIP-2 and combine it with a pre-trained Q-Former.
      * Gap bridging: A linear projection layer is used to bridge the gap between the visual encoder and the LLM.
      * Two-stage training: A two-stage training method is adopted. First, pre-training is performed on aligned image-text pairs to acquire visual language knowledge. Then, the pre-trained model is fine-tuned on a smaller but high-quality image-text dataset, and the designed dialogue templates are used to enhance the generation reliability and usability.

Next, we will briefly explain the specific implementation logic of LLM and the application of the ViT backbone structure in the model.

#### LLM and Vicuna

LLM is actually a large language model. AGI is actually Artificial General Intelligence. NLP can be roughly divided into two major tasks: NLP understanding tasks and NLP generation tasks. The difference between these two types of tasks is mainly reflected in the input and output forms.

The characteristic of comprehension tasks is that after inputting a sentence (article) or two sentences, the model finally determines which category they belong to. Therefore, they are essentially classification tasks, such as text classification, sentence relationship judgment, sentiment tendency judgment, etc.

The characteristic of generation tasks is that, given an input text, the model needs to generate a string of output text, such as chatbots, machine translation, text summarization, question-answering systems, etc.

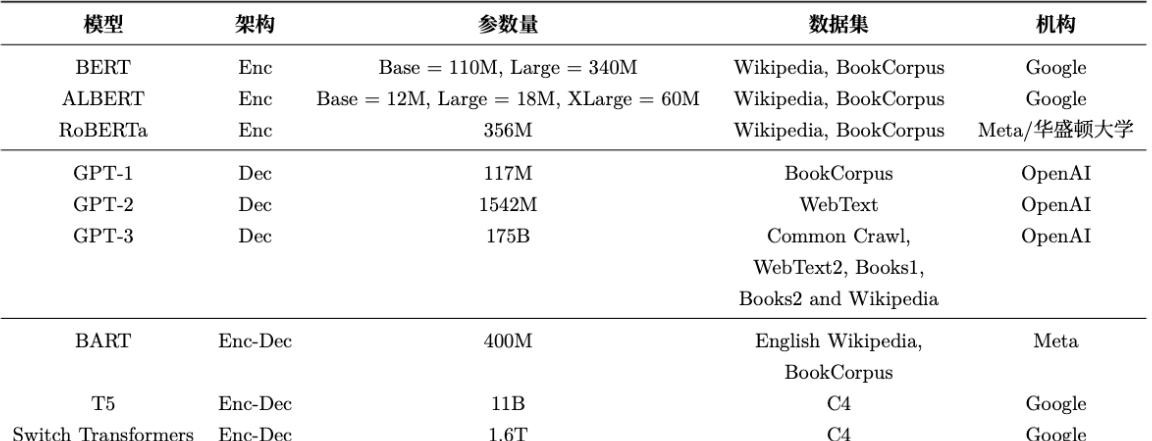
In fact, various NLP tasks have converged into two different pre-training model frameworks:

For NLP understanding tasks, the technical system has been unified into the "pre-training + fine-tuning" model represented by Bert;

For NLP generation tasks, the technical system has been unified into the "autoregressive language model (i.e., unidirectional language model from left to right)

+ Zero/Few Shot Prompt" mode represented by GPT.

At present, the development of transform-based LLM models has been very rapid. Most comprehension models are trained in this way. The following figure is a reference for the training parameters of some models.



#### VIT Backbone Structure

Dosovitskiy firstly applied the original Transformer model to image classification tasks and proposed ViT (Vision Transformer), a structure based entirely on the

self-attention mechanism. The authors believe that On the model dataset , without relying on CNN, Transformer can perform well in classification tasks. In order to convert images into sequence data that can be processed by the Transformer structure , the concept of image patches is introduced. First, the two-dimensional image is divided into blocks, and each image block is flattened into a one-dimensional vector. Then, each vector is linearly projected, and position encoding is introduced to add the position information of the sequence. In addition, a classification flag (class) is added before the input sequence data to better represent the global information. The ViT model is usually pre-trained on a large dataset and fine-tuned for smaller downstream tasks. On the ImageNet dataset, VIT-H/14 surpassed the EfficientNet model with an accuracy of 88.55% Top-1 , successfully breaking the monopoly of convolution-dominated networks on classification tasks, and is more efficient and scalable than traditional CNN networks . ViT is the first work of Transformer to replace standard convolution on large-scale datasets, laying an important foundation for the development of Transformer in computer vision tasks.



In the VIT algorithm, we split the image into patches and provide a sequence of linear embeddings of these patches as input to the Transformer. Image patches are

treated the same way as tokens (words) in NLP applications. We train the model for image classification in a supervised manner.

Transformer model lacks some of the inductive biases inherent in convolutional neural networks (CNNs), such as translation equivariance and locality, and therefore does not generalize well when training data is insufficient. Due to the lack of these inductive biases, the generalization ability of the Transformer model may be limited when it is trained with insufficient training data. Less training data may not be sufficient to capture complex image structures and contextual information, causing the model to perform poorly on unseen data. The position embeddings at initialization do not contain any information about the 2D position of the patch, and the spatial relationship between all patches needs to be learned from scratch. The MLP layer is local and translation equivariant , while the self-attention layer is global.

Therefore, ViT can achieve excellent results when pre-trained on a sufficient scale and transferred to tasks with less data. This is also an important reason why this system uses fewer high-quality datasets for training.

### modelscope

Modelscope is a collection of the most advanced open source models in the field of machine learning, providing developers with easy-to-use one-stop product services such as model building, training, and deployment. This section introduces some of the models we call and deploy using Modelscope , and explains their methods and principles.

#### Qwen

Qwen is a versatile language model family that includes models of various parameter sizes, such as Qwen (the base pre-trained language model, i.e., the base

model) and Qwen-Chat (the chat model, which is fine-tuned using human alignment techniques). The base model consistently shows excellent performance in many downstream tasks, while the chat model, especially the one trained using reinforcement learning with human feedback (RLHF), is very competitive. The chat model Qwen-Chat has advanced tool usage and planning capabilities, and Qwen-Chat can also show very competitive performance compared to larger models.

Qwen uses an improved version of the Transformer architecture. Specifically, it uses the training method of the recently open-source large language model LLaMA and makes the following improvements:

The embedding and output mapping do not share weights, thus achieving better performance at the expense of memory cost.

RoPE (Rotational Position Encoding) is used for position encoding. RoPE has been widely adopted in contemporary large language models, such as PaLM and LLaMA . In order to prioritize model performance and obtain higher accuracy, the inverse frequency matrix of FP32 accuracy is used instead of BF16 or FP16.

Pre-normalization (Pre-Norm) and RMSNorm are used for normalization. Pre-Norm is the most widely used method, and it has been shown to improve training stability compared to post-normalization. Recent research has proposed other methods to improve training stability, and the official said that they will be explored in future versions of the model. In addition, RMSNorm is used to replace the traditional layer normalization technique. This change improves efficiency without compromising performance.

SwiGLU is used as the activation function. It is a combination of Swish and Gated Linear Unit GLU. Preliminary experiments show that GLU-based activation functions generally outperform other baseline options such as GeLU . Following

common practice in previous studies, the dimension of the feed-forward network (FFN) is reduced from 4 times the hidden size to 8/3 of the hidden size.

Extrapolation capability

Qwen uses the following techniques to achieve context length expansion in the inference phase:

* NTK-aware interpolation, a training-free technique that adjusts the scale parameter to prevent loss of high-frequency information when extending the length.
* Dynamic NTK-aware interpolation, an improved version of NTK-aware interpolation, can dynamically change the scale parameter in blocks to avoid a significant drop in performance. These techniques enable Qwen to effectively extend the context length without affecting computational efficiency and accuracy.
* the dot product of Q and V according to the ratio of context length to training length , ensuring that the entropy of the attention value remains stable as the context length grows.
* Use layered window Self-Attention to limit attention to a context window to prevent the model from focusing on content that is too far away. Different window sizes are used at different layers, with lower layers using shorter windows and higher layers using longer windows. This is because the official observed that the Qwen model has different modeling capabilities at different levels when processing long contexts, and lower layers are more sensitive to the expansion of context length than higher layers. To this end, different window sizes are assigned to each layer, using shorter windows for lower layers and longer windows for higher layers.

Combining these technologies, the Qwen model can process long sequences of 8192 tokens in the inference stage, with excellent extrapolation capabilities.

Given qwen 's strong ability in natural language reasoning, we connected it to the subsequent output of MinigGPT-4 and processed the results a second time to generate the prompts required for the image-generated video and more detailed appreciation paragraphs.

#### Image quality damage analysis

Image quality damage analysis is done by modelscope , which uses the resnet50 structure and the 130,000 most popular images from the image website pexels as a training set. The model is trained to evaluate clarity, point noise, and compression noise by simulating various degradations .

This system calls this model to perform damage analysis on the pictures input by the user to determine the picture quality.

### Image Aesthetic Assessment

In this system, the image aesthetic evaluation algorithm is mainly implemented using two algorithms. One is the Image-reward algorithm, which calculates the human preference scoring algorithm. It was originally used to calculate the human preference of multimodal generated images, and is used here to evaluate the human preference of the scroll. The other is to use a simple neural network model to learn and train the image and evaluate the aesthetic value of the image.

#### Image-reward

ImageReward — The first general-purpose human-preference reward model for text-to-image, effectively encoding human preferences. Image Reward outperforms existing scoring models and metrics in terms of human evaluation, making it a

promising automatic metric for evaluating text-to-image generation. Building on this, we propose Reward Feedback Learning ( ReFL ), a direct conditioning algorithm for rater-optimized diffusion models. Both automatic and human evaluations support the advantages of ReFL over comparable methods. [ 3 ]

During the training process, it is difficult to promote it on a large scale due to the limited labor costs. The goal is to model human preferences based on annotations, so that virtual evaluators can get rid of their dependence on humans.

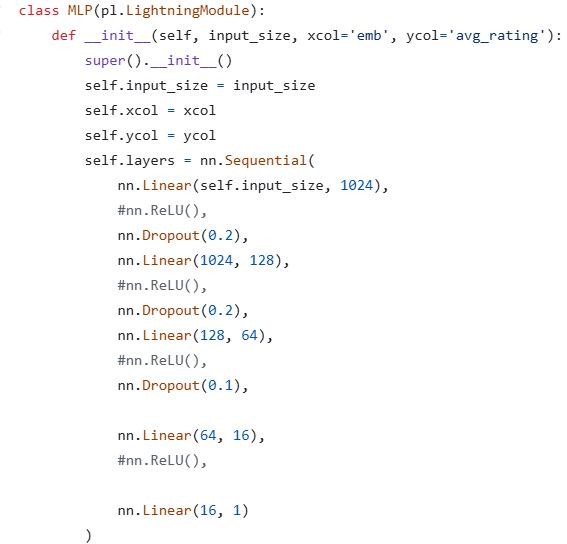
Similar to the training of the language model using the reward model, the preference is marked as a ranking. The loss function is as follows:



And, we use BLIP as the backbone of ImageReward because it outperforms traditional CLIP in preliminary experiments. We extract image and text features and combine them with cross attention and MLP to generate a scalar for preference comparison.

#### CLIP+MLP Aesthetic Score Predictor

In this section, we used a simple neural network model, which mainly used CLIP image feature extraction as input, wrote an MLP predictor, and implemented a simple aesthetic evaluation function.



### Image-to-video

#### ST -UNet

##### UNet Infrastructure

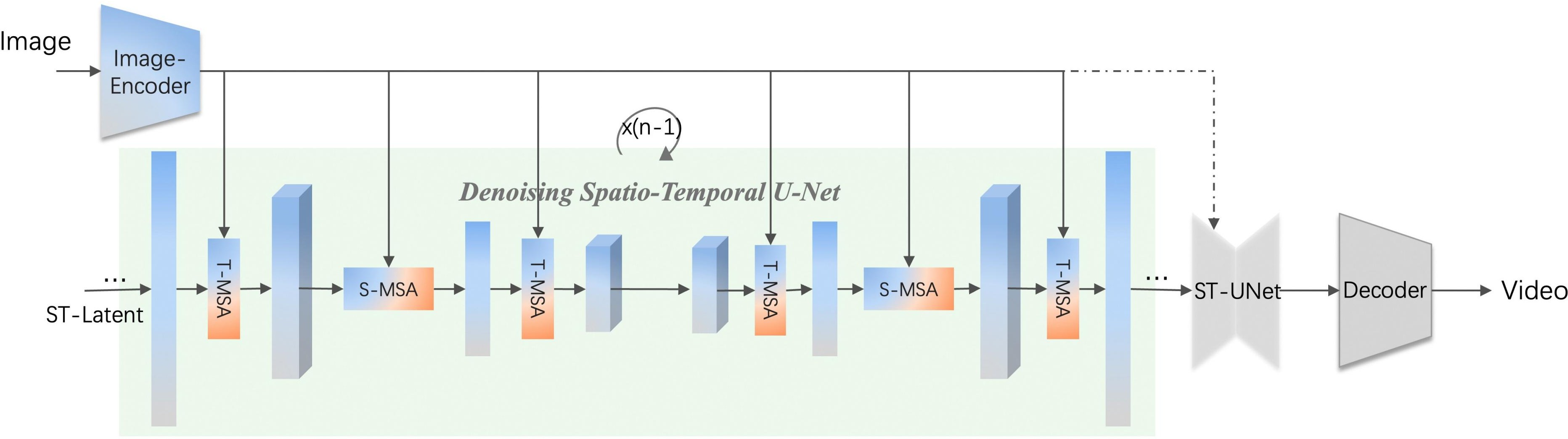
UNet is a neural network architecture commonly used for image segmentation. UNet consists of two parts: an encoder and a decoder, which are connected by skip connections.

* + **Encoder (downsampling path):** extracts high-level features of the input data, gradually reduces the spatial resolution through convolutional and pooling layers, and increases the level of abstraction of features.
  + **Decoder (upsampling path):** Restores the spatial resolution of the data, gradually restores the feature maps through upsampling and convolutional layers, and

combines the features of the corresponding layers of the encoder (skip connections) to preserve detail information.

##### Spatiotemporal modeling

Spatiotemporal UNet (ST- UNet ) is a model that extends 2D UNet to 3D UNet , specifically for spatiotemporal modeling. By introducing a temporal modeling layer, ST- UNet is able to process video data or other sequence data containing a temporal dimension. In order to take advantage of diffusion models (LDMs) pre-trained on large-scale image data, such as Stable Diffusion, these pre-trained models are integrated in ST- UNet , and a temporal processing block is introduced. The following are the specific implementation details of ST- UNet for spatiotemporal modeling in latent space:



ST- UNet inherits the structure of 3D UNet and extends it by introducing a modeling layer with a temporal dimension. The entire architecture consists of an encoder, a decoder, and jump connections, similar to traditional UNet , but processes in both spatial and temporal dimensions.

In a single UNet block, the following four basic building blocks are used:

* + **Spatial Convolution**
  + **Temporal Convolution**
  + **Spatial Transformer**
  + **Temporal Transformer**

1. **Spatial Convolution**

The spatial convolution block inherits from LDMs and is used to extract features in the spatial dimension. Each spatial convolution block contains two 3x3 convolution operations followed by a ReLU activation function.

1. **Temporal Convolution**

The temporal convolution block is newly introduced and is used to extract features in the time dimension. The specific implementation is as follows:

* + **Convolution operation:** Temporal convolution is performed using a 1×1×3 kernel. This design ensures that the temporal receptive field is large enough to capture temporal dependencies.
  + **Stacked convolution:** Four such convolutional layers are stacked together to fully capture changes in the temporal dimension.

1. **Space Transformer**

The spatial transformer block utilizes pre-trained features in LDMs to capture complex feature interactions in the spatial dimension. It usually contains a multi-head attention mechanism and a feed-forward network.

1. **Time Transformer**

The Temporal Transformer block is newly introduced to capture feature interactions in the temporal dimension. The specific implementation is as follows:

* + **Transformer layer:** Each Temporal Transformer block contains a standard Transformer layer including a multi-head attention mechanism and a feed-forward network.
  + **Stacked Transformers:** Multiple transformer layers are stacked together to enhance modeling capabilities in the time dimension.

1. **Encoder-Decoder Path**

The encoder path consists of multiple layers of spatial convolution, temporal convolution, spatial transformer, and temporal transformer blocks. Each layer reduces the resolution by downsampling (usually max pooling) and extracts progressively more abstract features.

The decoder path is symmetrical with the encoder path, and the resolution is gradually restored by upsampling (usually deconvolution or bilinear interpolation). Each layer includes spatial convolution, temporal convolution, spatial transformer, and temporal transformer blocks, and features are fused with the corresponding layer of the encoder path through skip connections.

1. **Skip Connections**

Skip connections are used to transfer feature maps between the encoder and decoder. Each upsampling layer concatenates the corresponding encoder layer feature map to the decoder layer to preserve high-resolution detail information.

1. **Final Output**

In the last layer, a 1x1x1 convolution kernel is used to convolve the feature map to the required number of output channels. Depending on the task requirements, an appropriate activation function is used, such as Softmax (for multi-class segmentation) or Sigmoid (for binary segmentation).

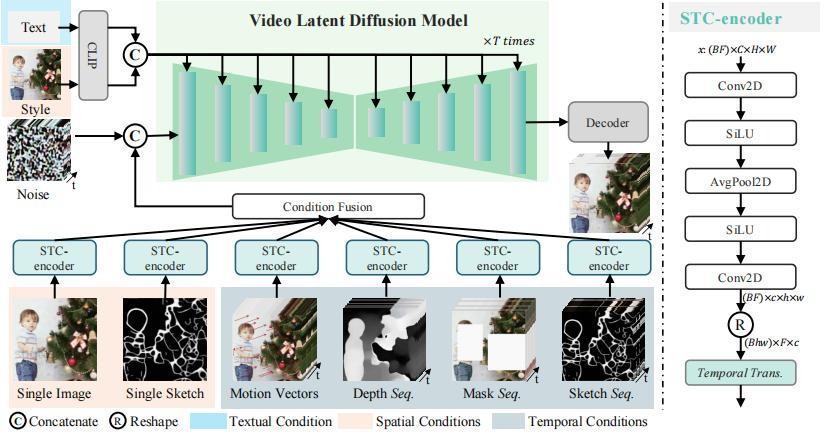
##### Advantages

* + **Efficient spatiotemporal modeling:** By combining spatial and temporal convolutions and transformer blocks, ST- UNet is able to efficiently capture complex patterns in spatiotemporal data.
  + **Utilizing pre-trained models:** By introducing spatial convolutions and transformer blocks of LDMs, ST- UNet is able to utilize features pre-trained on large-scale image data to improve model performance.
  + **Preserving detail information:** The skip connection mechanism ensures that high-resolution detail information is preserved in the decoder, improving segmentation accuracy.

#### VLDM

VLDM is a technique that applies diffusion models to video data. By modeling in latent space, VLDM can effectively process and generate video data. The method mainly includes three core components: encoder, diffusion model and decoder, and introduces a temporal modeling layer to enhance the ability to capture spatiotemporal features. Video data usually contains noise and low resolution, which affects the visual quality and consistency of the video. VLDM not only improves the video

resolution through the denoising process, but also maintains consistency in time and space.



##### Encoder: From raw video to latent space representation

The task of the encoder is to convert the raw video data into a latent space representation in preparation for the subsequent denoising process.

* + **3D Convolutional Network** : We use 3D convolutional networks to extract spatiotemporal features from video data. 3D convolutions operate simultaneously in both temporal and spatial dimensions, and can capture dependencies between video frames . Each convolutional layer is followed by a nonlinear activation function (such as ReLU ) to improve feature expression capabilities.
  + **Dimensionality reduction** : To reduce computational complexity, we add downsampling operations (such as pooling layers) after the 3D convolutional network to project high-dimensional video data into a low-dimensional latent space. This step not only reduces the amount of data, but also retains the main features of the video.

##### Forward Diffusion Process: From Latent Space Representation to Noise Data

Latent space representation into data close to Gaussian distribution by gradually adding noise, providing a training target for the backward diffusion process.

* + **Gradually add noise** : Add noise in multiple time steps, adding a certain amount of Gaussian noise to the latent space representation at each step. This process simulates the situation where the data is gradually contaminated by noise, allowing the model to learn how to recover the original data from different degrees of noise.
  + **Target distribution** : By gradually adding noise, the latent space representation eventually approaches a Gaussian distribution. The goal of this step is to map complex video data to a standard probability distribution, making the back diffusion process easier to model and train.

##### Backward Diffusion Process: From Noisy Data to Clean Latent Space Representation

The back-diffusion process recovers the clean video representation from the latent space through a denoising operation while improving the resolution.

* + **Denoising model** : We designed a neural network model to gradually remove noise and recover a clean latent space representation. At each time step, the model predicts and subtracts the noise based on the current noisy data, so that the data gradually returns to a noise-free state.
  + **Resolution enhancement** : While denoising , the model also performs

super-resolution processing. By introducing a resolution enhancement module (such as a deconvolution layer) in each denoising step , the details and clarity of the video

are gradually restored and improved, making the generated video not only noise-free but also with higher resolution.

##### Temporal Modeling Layer: Enhancing Temporal Consistency

Enforces temporal consistency by capturing the temporal dependencies between frames in the video.

* + **Temporal convolution** : We use 1x1x3 convolution kernels for temporal convolution operations in the latent space representation. Temporal convolution can capture short-term dependencies between adjacent time frames , and by stacking multiple temporal convolution layers, the model can learn more complex temporal patterns.
  + **Temporal Transformer** : To capture dependencies over a longer time frame, we introduce the Temporal Transformer. The Temporal Transformer consists of a

multi-head attention mechanism and a feed-forward network, which can integrate information and interact features over a long time span, ensuring the temporal coherence and consistency of the video.

##### Decoder: From latent space representation to high-resolution video

The decoder converts the processed latent space representation back to high-resolution original video data.

* + **3D transposed convolutional network** : We use 3D transposed convolutional network to gradually increase the resolution of the data. 3D transposed convolution can effectively restore the spatial and temporal resolution of the video, and the latent

space representation is enlarged back to the size of the original video through deconvolution operation.

* + **Upsampling operations** : To ensure that the generated video has the same resolution as the input video, we introduce upsampling operations (such as bilinear interpolation) during the decoding process. These operations further enhance the resolution and details of the video while maintaining spatiotemporal consistency.

##### Spatial consistency enhancement: Maintaining the spatial coherence of the video

In the process of restoring high-resolution videos, spatial consistency is maintained through spatial convolution and spatial transformer layers.

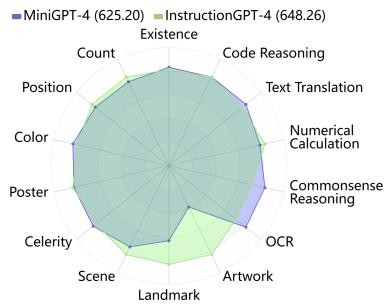
* + **Spatial convolution** : We use spatial convolution operations in the decoder to perform convolution in the spatial dimension. Spatial convolution can effectively process the spatial features within the video frame and ensure the spatial consistency and coherence of each frame of the video.
  + **Spatial Transformer** : To capture complex spatial feature interactions, we introduce the spatial transformer. The spatial transformer integrates and interacts features in each frame of the video through a multi-head attention mechanism and a feedforward network, ensuring the spatial consistency and high quality of the generated video.

### Fine-tuning and Dataset Cleansing Approach from InstructionGPT-4 for Mini-GPT4[4]

When building and training multi-modal large language models, the quality of the dataset is crucial to the model's performance. To enhance the performance of MiniGPT-4 and make it more suitable for our applications, we particularly focused on and optimized the dataset for fine-tuning. Our dataset cleaning and optimization techniques mainly include the following aspects:

Referring to InstructionGPT-4, which proposes a series of metrics to evaluate the quality of multi-modal instruction data, we have introduced an effective and trainable data selector to automatically identify and filter low-quality visual-linguistic data. By adopting this approach, InstructionGPT-4 outperforms the original MiniGPT-4 in various evaluations.

Commonly used instruction-tuned datasets contain a large number of low-quality instances with incorrect or irrelevant responses, which may negatively impact model performance. Research has shown that fewer but higher-quality instruction-tuned data can effectively enable multi-modal large language models to produce better outputs.



Comparison of MME evaluation (InstructionGPT-4 vs. MiniGPT-4)

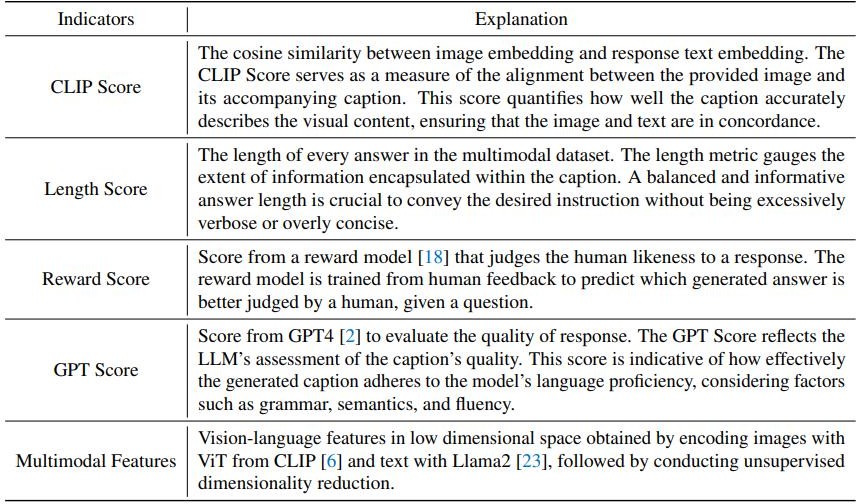
Multimodal large language models are typically trained in two stages: firstly, pre-training on image-text pairs, and then fine-tuning using supervised

visual-linguistic instruction data.

**Objective:** To propose a robust and effective data selector that automatically identifies and filters low-quality visual-linguistic data in existing datasets, ensuring that the model is trained on the most relevant and informative samples.

**Challenge:** There is currently a lack of comprehensive methods to evaluate the quality of visual-linguistic data.

Therefore, we have introduced several new metrics to evaluate the quality of multimodal instructional data, including CLIP scores, GPT scores, reward scores, length scores, and multimodal features for each visual-linguistic data.



To investigate the relationship between metrics and actual instruction data quality, we first segmented a series of different subsets from the original fine-tuning

data and recorded the performance of each fine-tuned model on the validation set as a label for data quality. Then, we calculated the metrics and multimodal data features for each subset and combined them into an embedding for each subset.

##### Self-attention network as a data selector:

After determining the data quality labels and embeddings, a self-attention network is applied as a data selector to determine the relationship between the true quality labels and embeddings.

This data selector can help identify which data subsets have higher quality, which can then be used to further optimize the model.

##### Spectral clustering:

To ensure the diversity of data distribution, spectral clustering was performed on the original 3.4K data used to fine-tune MiniGPT-4.

Spectral clustering is a graph-based clustering algorithm that groups data points into different clusters based on their similarities.

##### Quality label prediction and ranking:

After spectral clustering, the data selector is applied to each cluster to predict its quality label and rank them accordingly.

This ensures that higher-quality data subsets are used during fine-tuning, thereby improving the model's performance.

##### Model fine-tuning:

After selecting high-quality data subsets, InstructionGPT-4 is fine-tuned using the same training configuration as MiniGPT-4.

This fine-tuning process aims to leverage high-quality data to improve the model's instruction-following ability, thereby enhancing its performance in practical applications.

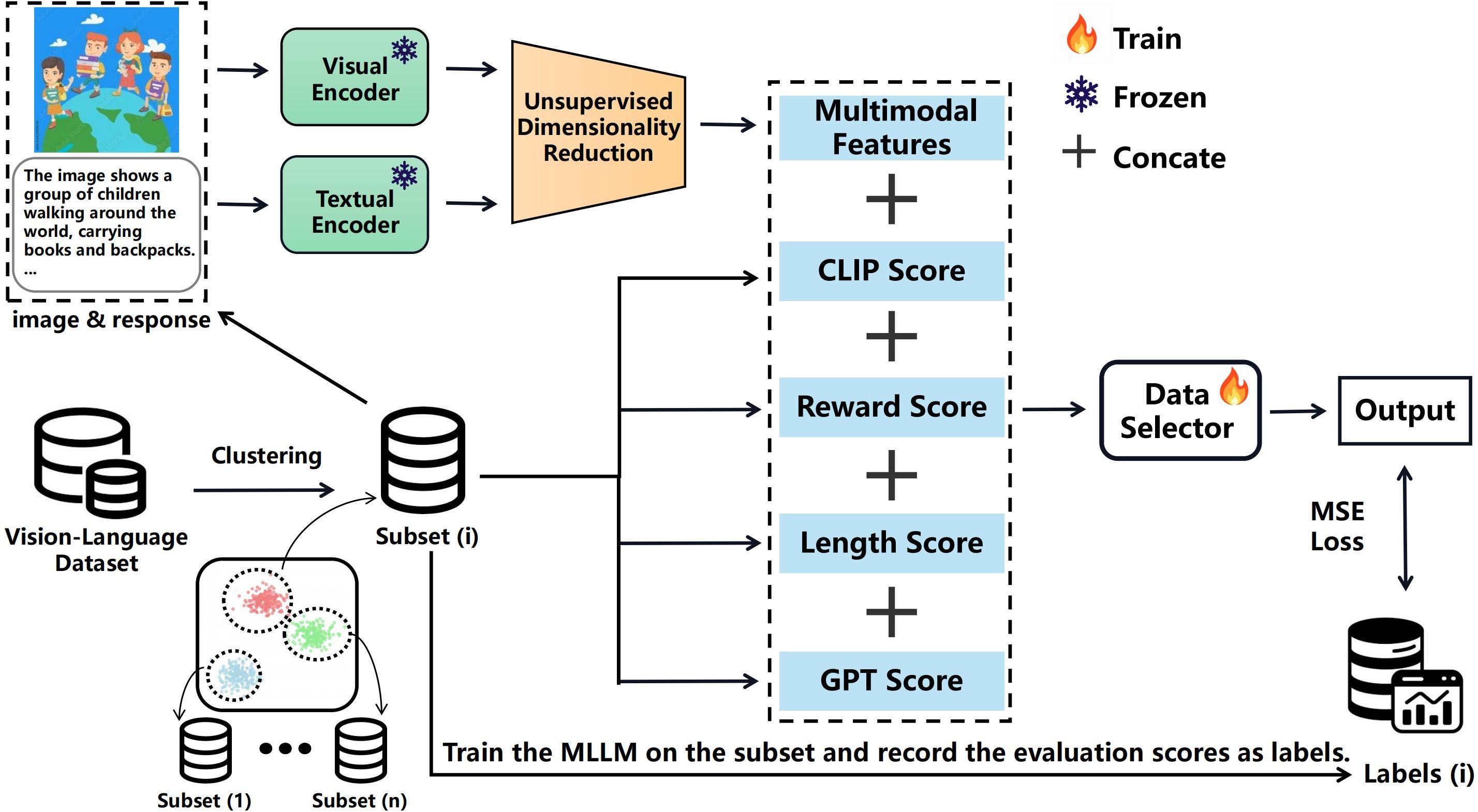


Figure: Overall procedures of the data selector. We first split the vision-language dataset into n subsets. Subsequently, we train the MLLM on each subset and record the evaluation scores as genuine quality labels. Additionally, we concatenate various indicators generated from these subsets to form embeddings. These embeddings were then used to train the data selector, with the objective of aligning the embeddings with the quality labels.

**Training:** Given a vision-language instruction dataset, a reasonable and straightforward strategy to obtain the genuine quality labels is to divide the original multimodal dataset into n subsets of equal size through clustering (e.g., K-means++).

For each subset i, we now obtain the embedding of every triplet in subset i, along with the quality label yi . We concatenate these embeddings into a single composite embedding denoted as ei , paired with its corresponding quality label yi . Having gathered a collection of such pairs (ei , yi) from all n subsets, we can then proceed to learn a data selector F to fit the embeddings {ei} n i=1 to quality labels {yi} n i=1.

The data selector could take various forms, such as a linear layer, an MLP, or a self-attention.

**Testing:**Given a multimodal dataset D of triplets x = (image, instruction, answer) with x ∈ D and a pre-trained MLLM (e.g., MiniGPT-4), our ultimate objective is to identify a high-quality subset S ⊂ D that, when utilized for fine-tuning, leads to the

improvement of the pre-trained MLLM. In order to select S from D and ensure its diversity, we first use a clustering algorithm (e.g., spectral clustering) to separate the images in D into K groups. The clustering algorithm is supposed to be different from the previous one because each of the former clusters shares the same quality label.

Suppose that the total amount of D is |D| and the i-th group’s amount is |Di |. We set

|S| = α as the size of the target subset.

For each x in D, we gain an embedding e(x) in Equation equation 1. We sort x according to the predicted label F(e(x)) and select Si from each group Di . Each Si contains top |Si | triplets x based on F(e(x)) from Di , i.e.,



At last, we combine these K subgroups:

S = S1 ∪ S2 ∪ . . . ∪ SK,

where S is the final high-quality dataset selected by the data selector.

##### evaluation and analysis:

Its evaluations focus on a wide range of complex open-domain multimodal large language model benchmarks, including MME, MMBench, VQA datasets from LVLM-eHub, etc. Through rigorous experimentation, they demonstrate that 200 pieces of data used for fine-tuning, which is 6% of the original scale, are enough to help InstructionGPT-4 achieve comprehensive superiority over MiniGPT-4 across these diverse multimodal tasks, with a +23 score enhancement on MME, a +1.55 score improvement on MMBench, and a +1.76% increase in performance on VQA datasets compared to MiniGPT-4. Specifically, InstructionGPT-4 outperforms MiniGPT-4 in 8 out of 14 tasks within MME, 13 out of 20 abilities in MMBench, and

excels in all four VQA datasets included in LVLM-eHub. It demonstrates that the data quality in vision-language instruction tuning can outweigh the quantity. In addition, this shift towards prioritizing data quality presents a new and more efficient paradigm that can generally improve the fine-tuning stage of MLLMs.

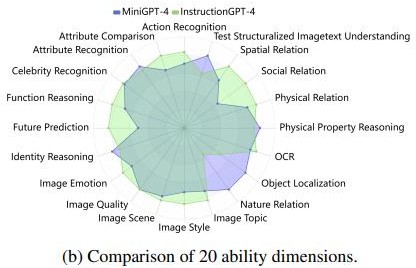
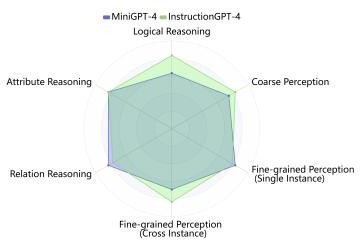


Figure: Comparison of MMBench evaluation (InstructionGPT-4 vs. MiniGPT-4).

### Experiment contents

### Dataset Processing Algorithm

#### DataSelector.py algorithm idea

Call CLIP and ImageReward models to calculate the image-text correspondence score and image aesthetic score, normalize them, and simply combine the two scores as the final criterion for selecting high-quality data, with weights CLIP0.8 and ImageReward0.2. Read the image path and corresponding style category of the initial dataset (i.e. the second stage targeted training dataset of MiniGPT-4, our landscape painting dataset). The 30 data with the highest scores in each style are used as the final high-quality dataset.

//Dataset reading example dataset\_dir = ' Style\_classification '

images\_by\_class = defaultdict ( list ) //Initialize a dictionary to store pictures. Use collections.defaultdict to initialize a dictionary where the key is the class name and the value is a tuple list of picture ID and picture path.

for filename in os . listdir ( dataset\_dir ): if filename . endswith ( '.jpg' ):

// Check if the file name ends with .jpg and construct the corresponding .json file name for it

image\_path = os . path . join ( dataset\_dir , filename ) json\_filename = filename . replace ( '.jpg' , ' .json ' ) json\_path = os . path . join ( dataset\_dir , json\_filename )

//Try to open the corresponding .json file and read the data in it. If successful, extract the name of the first tag, which is the class name . If the .json file does not exist, it means that the data is invalid and skip the current loop.

try :

with open ( json\_path , 'r' ) as json\_file : data = json . load ( json\_file )

if 'labels' in data and data [ 'labels' ]: label\_name = data [ 'labels' ][ 0 ][ 'name' ]

image\_id = os . path . splitext ( filename ) [ 0 ]

images\_by\_class [ label\_name ]. append (( image\_id ,

image\_path ))

except FileNotFoundError :

*# Ignore images that do not have a corresponding .json file*

continue

//Data set filtering selected\_images = {}

for class\_name , image\_list in images\_by\_ class . items (): scores = {}

for image\_id , image\_path in image\_list :

clip\_score = evaluation\_function1 ( image\_path ) image\_reward\_score = evaluation\_function2 ( image\_path )

//Do normalization

normalized\_clip\_score = normalize ( clip\_score , 1 , 10 ) normalized\_image\_reward\_score = normalize ( image\_reward\_score , 1 , 10 )

weighted\_score = ( normalized\_clip\_score \* 0.8 ) + ( normalized\_image\_reward\_score \* 0.2 )

scores [ image\_id ] = weighted\_score

// Sort the images in each category by score from high to low, and select the image paths corresponding to the first 30 IDs

sorted\_ids = sorted ( scores , *key* = scores . get , *reverse* = True )[: 30 ] selected\_images [ class\_name ] = [ image\_path for image\_id , image\_path in

image\_list if image\_id in sorted\_ids ]

Filter results:



#### StyleClassification.py algorithm ideas

1. Data preprocessing and feature selection: color histogram, gray-level co-occurrence matrix

IMAGE\_SIZE = ( 224 , 224 ) # Scaling to uniform size CROP\_SIZE = None # Crop size is not enabled

# Loading images and labels is similar to the previous dataset reading, so it is omitted here

# Feature extraction and dataset preparation

features = [] labels = [] label\_names = []

//Preprocessing

for label\_name , image\_paths in images\_by\_ class . items (): for image\_path in image\_paths :

image = cv2 . imread ( image\_path ) if image is not None :

image = cv2.resize ( image , IMAGE\_SIZE ) if CROP\_SIZE is not None :

h , w = image . shape [: 2 ]

start\_x = ( w - CROP\_SIZE [ 1 ] ) // 2 start\_y = ( h - CROP\_SIZE [ 0 ] ) // 2

image = image [ start\_ y : start \_y + CROP\_SIZE [ 0 ], start\_x : start\_x

+ CROP\_SIZE [ 1 ]]

//Calculate HSV color histogram : Convert the image from BGR color space to HSV color space. Then, calculate the color histograms of H, S, and V channels respectively and merge them into an array hist\_color

hsv = cv2 . cvtColor ( image , cv2 . COLOR\_BGR2HSV ) h , s , v = cv2 . split ( hsv )

num\_bins = 32 # Increase the number of bins to improve resolution hist\_h = cv2.calcHist ([ h ], [ 0 ], None , [ num\_bins ], [ 0 , 180 ] ) hist\_h = hist\_ h . flatten ()

hist\_s = cv2.calcHist ([ s ], [ 0 ], None , [ num\_bins ], [ 0 , 256 ] ) hist\_s = hist\_s.flatten ( )

hist\_v = cv2.calcHist ([ v ], [ 0 ], None , [ num\_bins ], [ 0 , 256 ] ) hist\_v = hist\_ v . flatten ()

#Merge the color histogram of H, S, and V channels hist\_color = np . concatenate (( hist\_h , hist\_s , hist\_v ))

//Convert the image to grayscale. Then, use the graycomatrix function to calculate the gray-level co-occurrence matrix (GLCM). Next, extract the four features of contrast, correlation, energy, and homogeneity from the GLCM and store them in an array glcm\_features

# Calculate the characteristics of the gray-level co-occurrence matrix gray = cv2 . cvtColor ( image , cv2 . COLOR\_BGR2GRAY )

glcm = graycomatrix ( gray , [ 1 ], [ 0 ], 256 , symmetric = True , normed =

True )

# Calculate contrast, correlation, energy and homogeneity separately contrast = graycoprops ( glcm , 'contrast' )[ 0 , 0 ]

correlation = graycoprops ( glcm , 'correlation' )[ 0 , 0 ]

能源 = graycoprops ( glcm , 'energy' ) [ 0 , 0 ]

Homogeneity = graycoprops ( glcm , 'homogeneity' ) [ 0 , 0 ] # Put these features into an array

glcm\_features = np . array ([ contrast , correlation , energy ,

homogeneity ])

// Merge the color histogram and GLCM features into a feature vector feature\_vector and normalize it using the normalize function

# Merge features

feature\_vector = np . concatenate (( hist\_color , glcm\_features )) feature\_vector = normalize ( feature\_vector . reshape ( 1 , - 1 ), axis = 1 ).

flatten ()

# Add to feature list and label list features . append ( feature\_vector ) labels . append ( label\_name )

# Update the tag name list

if label\_name not in label\_names : label\_ names . append ( label\_name )

# Update the list of label names to the encoder outside the loop label\_encoder = { name : idx for idx , name in enumerate ( label\_names )} encoded\_labels = [ label\_encoder [ label ] for label in labels ]

# Convert to NumPy array features = np . array ( features )

encoded\_labels = np . array ( encoded\_labels ) # Divide the dataset

X\_train , X\_test , y\_train , y\_test = train\_test\_ split ( features , encoded\_labels , test\_size = 0.2 , random\_state = 42 )

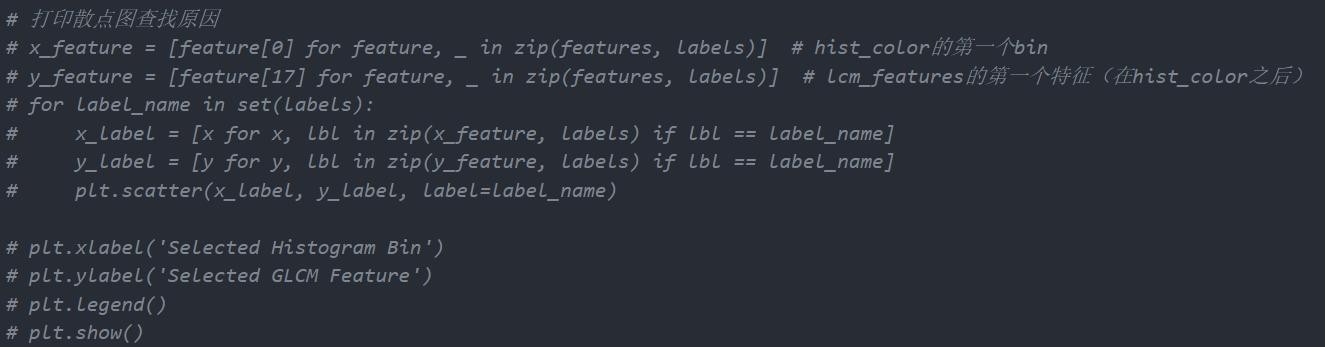
Tuning process:

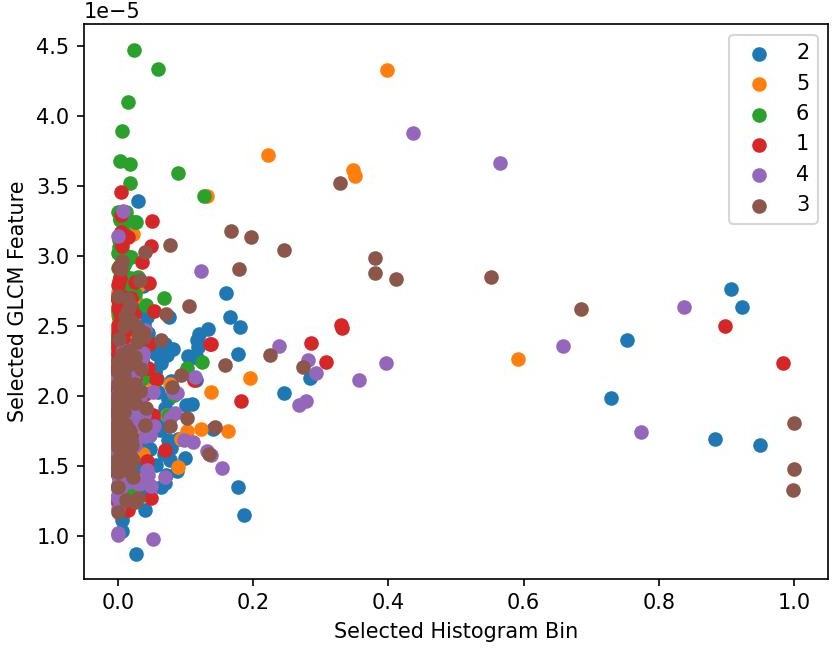
The initial feature selection was image tone and texture. After preliminary testing with

the SVM classification algorithm, the accuracy was only 0.3610. Even after parameter

tuning, it was only 0.4108. This may be due to improper selection of principal

components or poor quality of the data set itself, or the SVM classification algorithm is not suitable for handling such problems. Print a scatter plot to find out the reason:



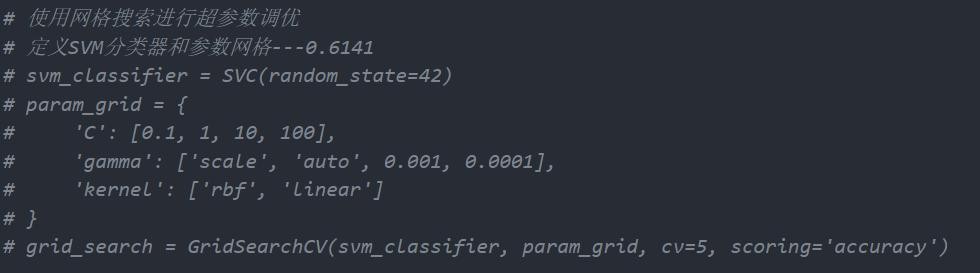


You can see that the Selected Histogram Bin is not distinguished. The previous code only calculated the histogram of the H (hue) channel in the HSV color space. Consider calculating the histograms of the S (saturation) and V (value) channels at the same time and merging them into the feature vector. The accuracy is improved to 0.6141.

1. Classifier training

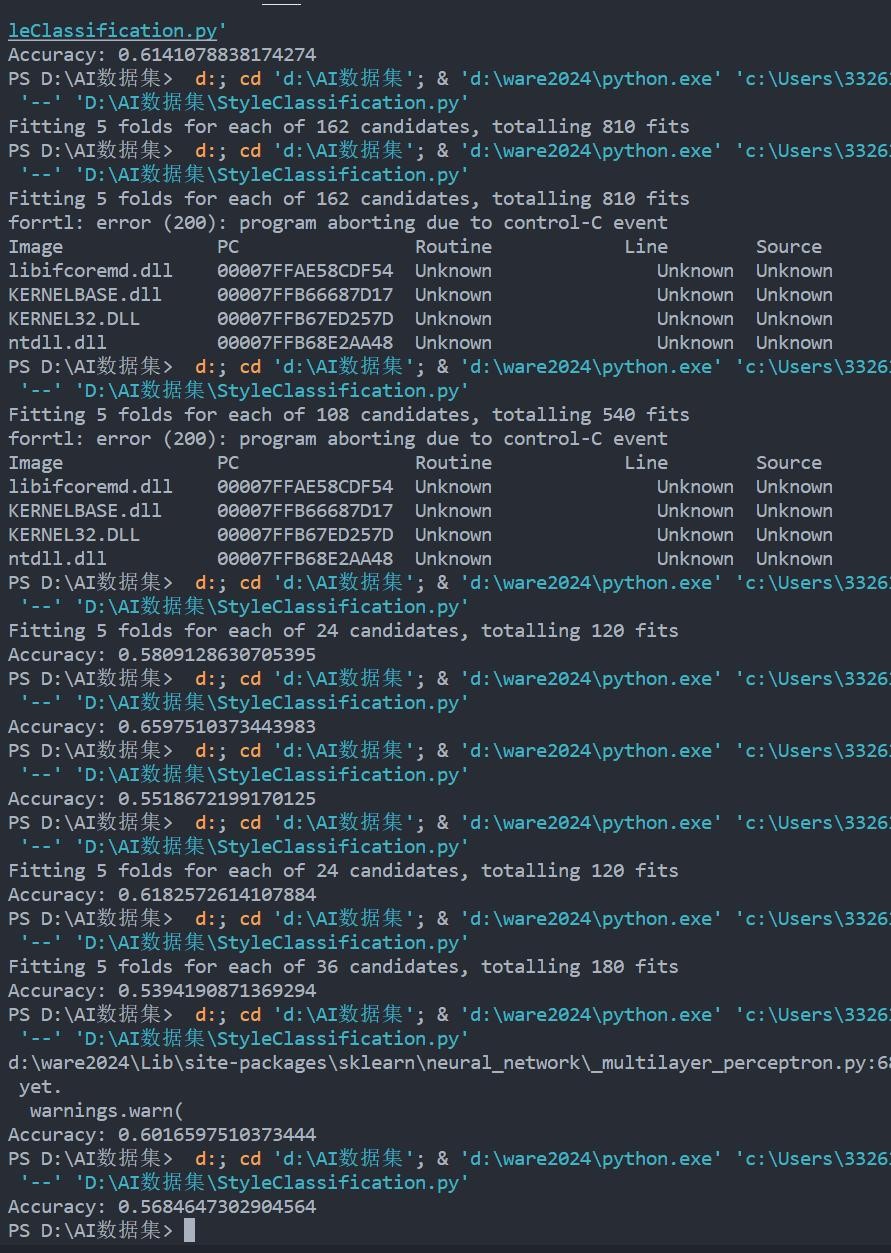
So far, we have completed the preprocessing and feature extraction of the data set, and converted these features and corresponding labels into NumPy arrays to divide the data set. Next, we can call the classifier supported by Python to train the model.

Parameter tuning using GridSearchCV



We tried different classification algorithms and finally selected the random forest algorithm, which had an accuracy of 0.6597.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classification  Algorithms | SVM | Logistic  Regression | Random  Forest | Gradient  Boosting | K nearest  neighbors | Decision  Tree |
| Accuracy | 0.6141 | 0.5809 | 0.6597 | 0.5518 | 0.6182 | 0.5394 |



This part needs further exploration and optimization. In addition, this homologous dataset has relevant model training to achieve style transfer, and the results are as follows:





Maybe it can be used as a future expansion of our project

#### Watermark Removal

Our video-text set has very accurate descriptions and rich video content, but it has watermarks, which will inevitably affect the training effect.

PaddleOCR is an OCR (Optical Character Recognition) system developed based on PaddlePaddle. Its technical system includes modules such as text detection, text recognition, text direction detection and image processing.

Use PaddleOCR to recognize text in the video. Because there may be other text in the video, we need to determine whether it is the watermark part. We use IOU and preset threshold. If it is greater than the gate value, it means that there are more overlapping parts and it can be considered as a watermark.

Part of the code:





#### Image-Video Collection Cleaning

5000 data from the WebVid dataset were selected as the video set. The video set needs to be cleaned before pairing with pictures. In the SVD paper, the author calculated four scores between each video-text pair from four angles, namely: CLIP score , OCR detection score , optic flow score , and aesthetic score .

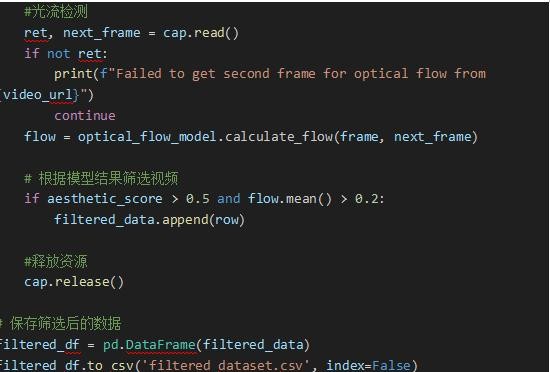
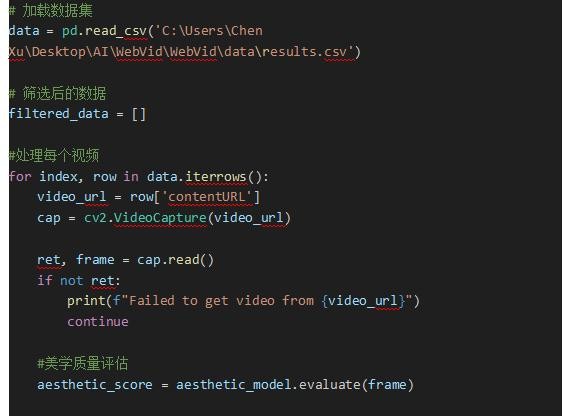
In this dataset, the paper was referenced and the following two scoring criteria were selected to clean the video set:

* + - * Optic flow score: detects the optical flow. The greater the change between two frames, the higher the optical flow score will be. It is used to detect the size of the

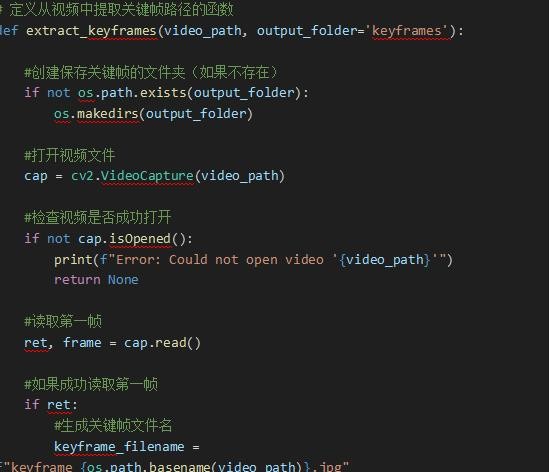
motion change of the video.

* + - * Aesthetic score: judge the aesthetics of the video itself.

Set the thresholds according to the actual situation and the paper reference, aesthetic score: 0.5, optic flow score: 0.2



After cleaning, use OpenCV to extract the first frame of the video from the video set to get its corresponding image, and write it into a new CSV file to form a video-image pair dataset.

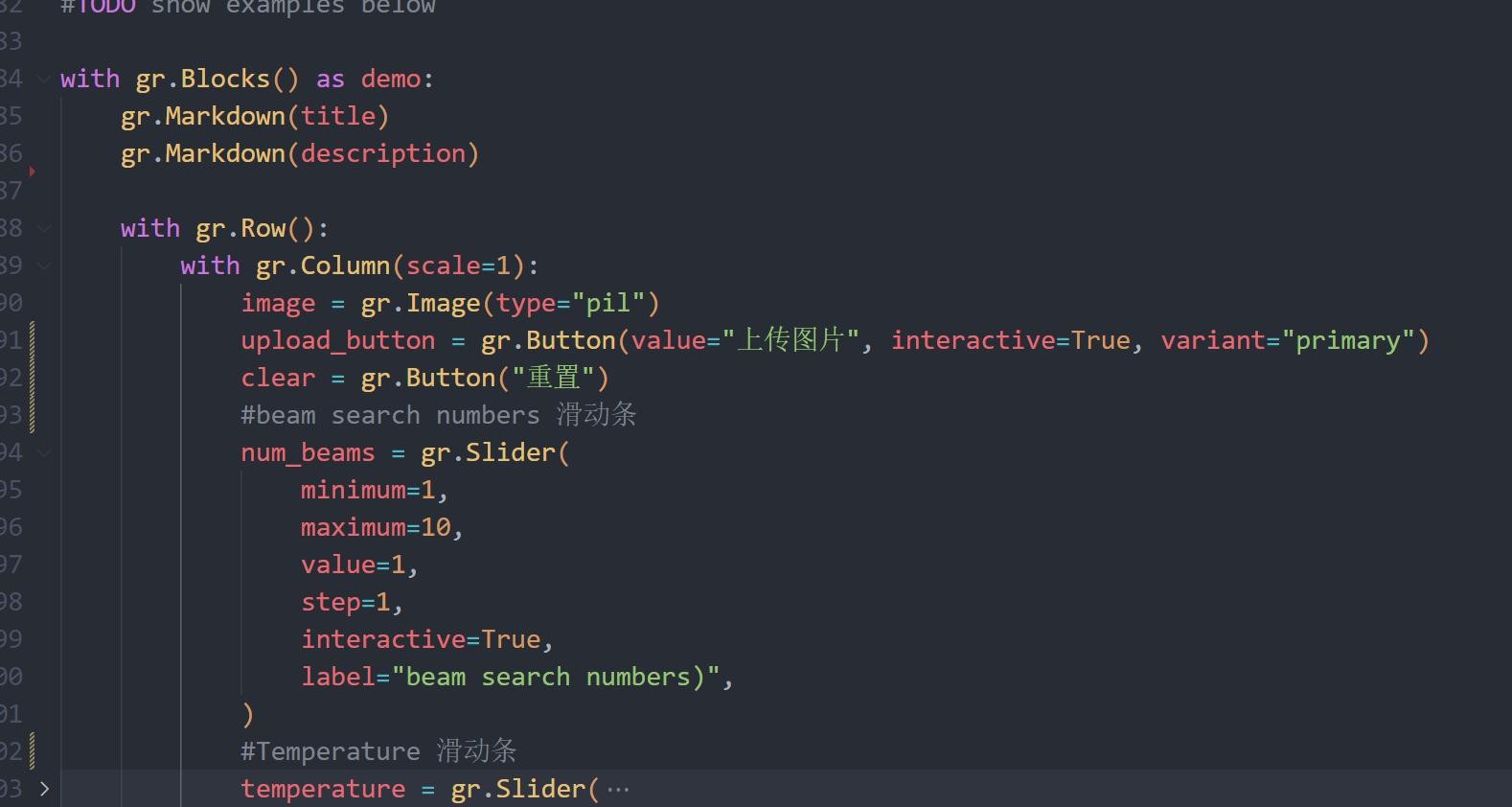


### Front-end and back-end connectivity calling algorithm

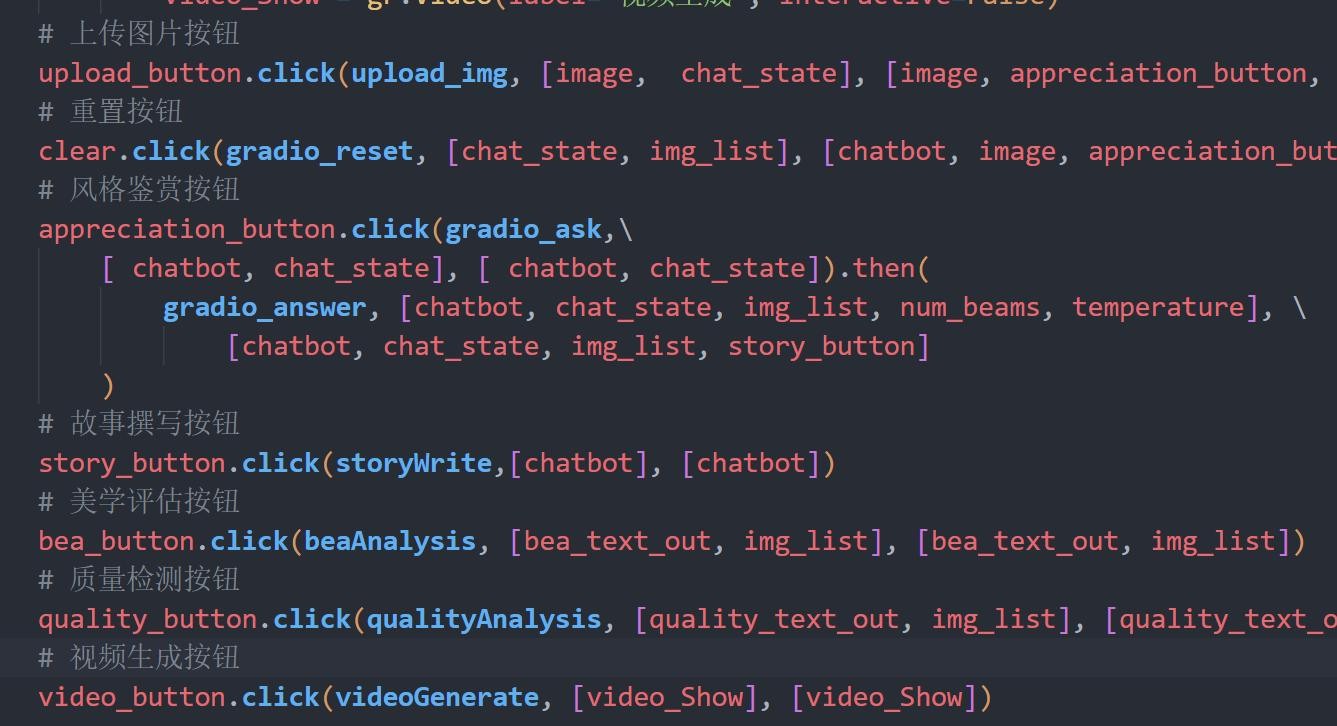
We used the Gradio library to create a feature-rich front-end interface designed to provide a user-friendly experience. Through this interface, users can easily upload pictures and set relevant parameters as needed . After the picture is uploaded, the function button on the right will automatically unlock, providing users with a wide range of operation options. Users can appreciate the style and analyze the content of the picture to get its aesthetic analysis. At the same time, users can also use the story writing function to add narrative to the style appreciation of the picture. After completing the style appreciation and story writing, users can continue to generate videos at the bottom of the interface. By clicking the corresponding button, the system will automatically process and generate a dynamic video based on the uploaded picture, showing the changes in the picture in a smooth animation form. And users can also perform aesthetic evaluation and quality inspection to judge the aesthetic level of the painting. Whether it is style construction, story writing or video generation, users

can easily appreciate, analyze and evaluate Chinese style paintings on this interface. We perform automatic layout of front-end components in the gr.row () function, and perform component logic binding externally, linking the components with the relevant logic functions written, so changing the component logic will not change the

front-end layout.



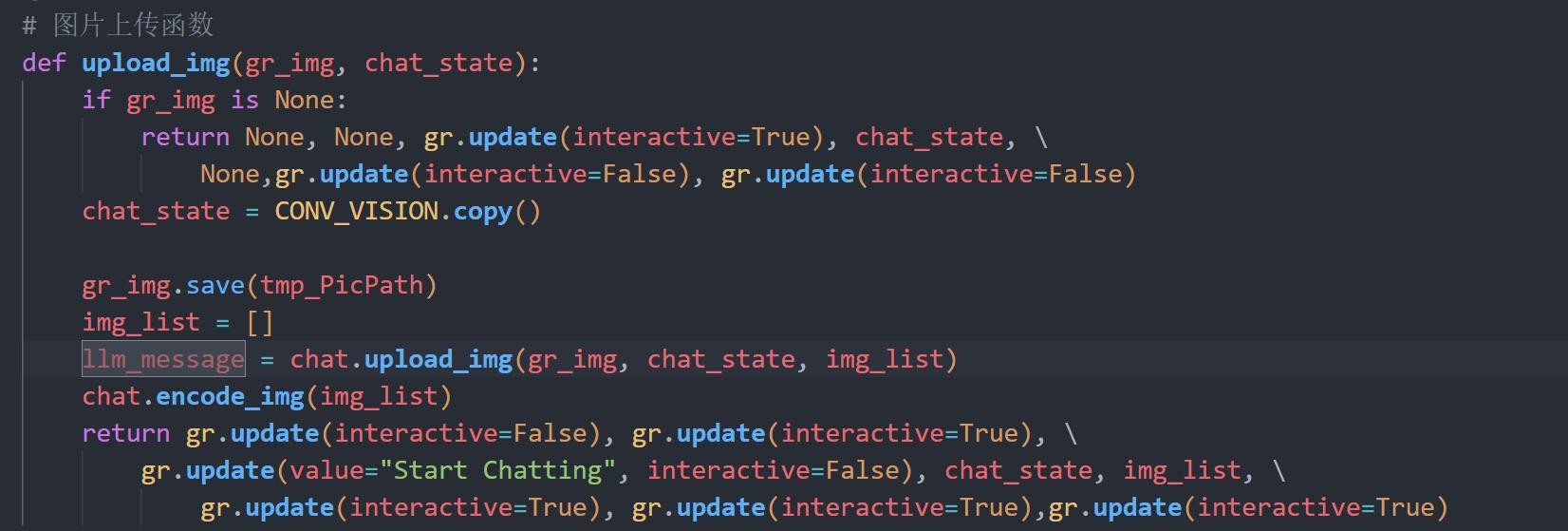
component layout



component Links

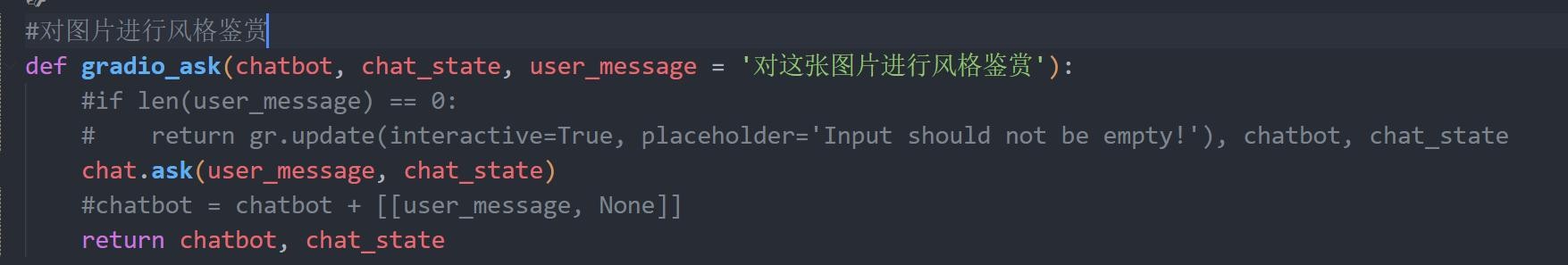


gradio\_reset , reset function, after clicking, clears the image cache and resets all components to facilitate users to use multiple times.

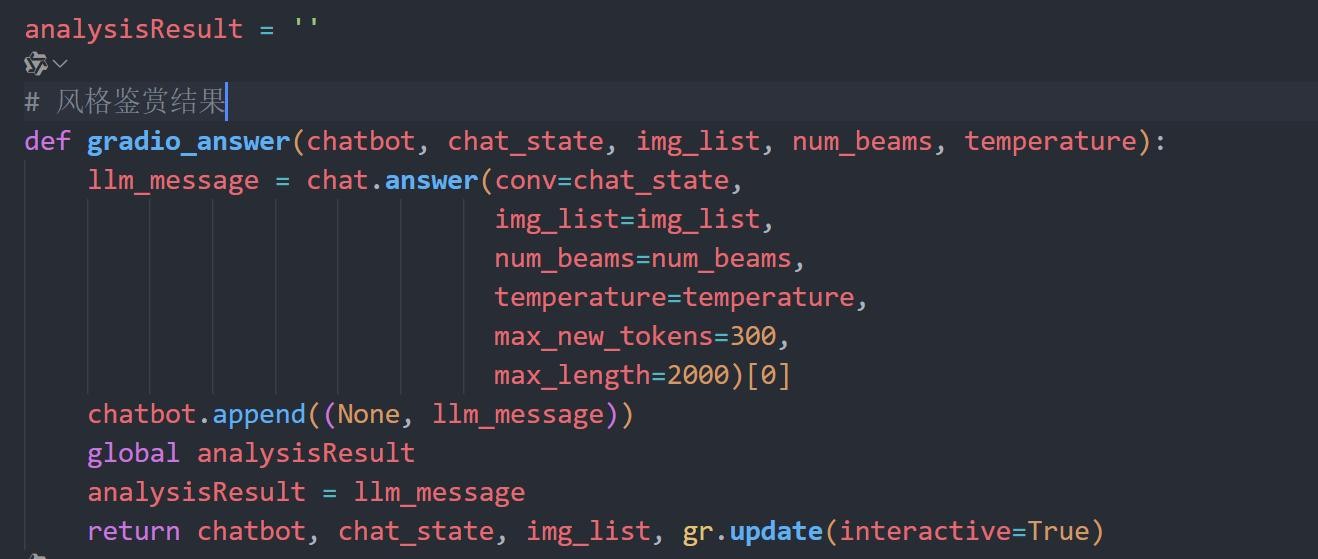


upload\_img, a picture upload function. After clicking on the picture to upload, the picture will be saved to a specified temporary path for easy subsequent calls . After the upload is successful, the remaining button components will be set to be interactive

to facilitate the user to proceed to the next step.



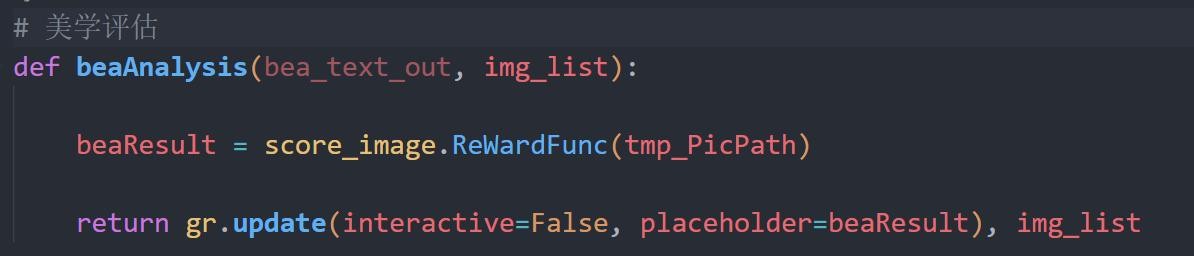
gradio\_ask , style appreciation button function, when pressed, calls the backend miniGPT model to perform style appreciation on the image.



gradio\_answer , style appreciation result function, after style appreciation, obtains the appreciation result of miniGPT backend, displays it on the front-end interface, and saves it to global variables for subsequent use.



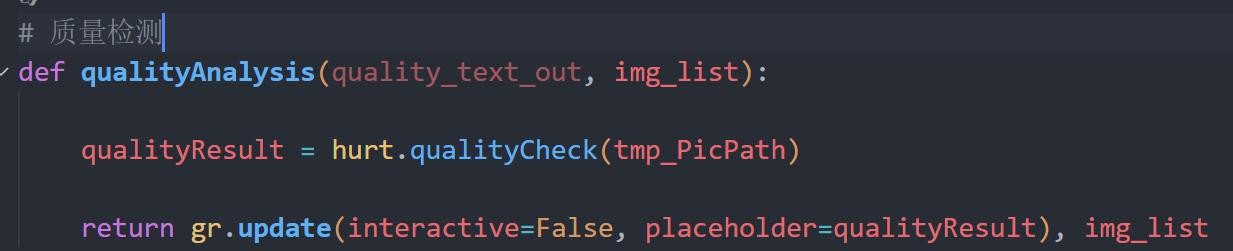
storyWrite , a story writing function, calls the Tongyi Qianwen model in the backend to write the story and returns the result to the frontend interface.



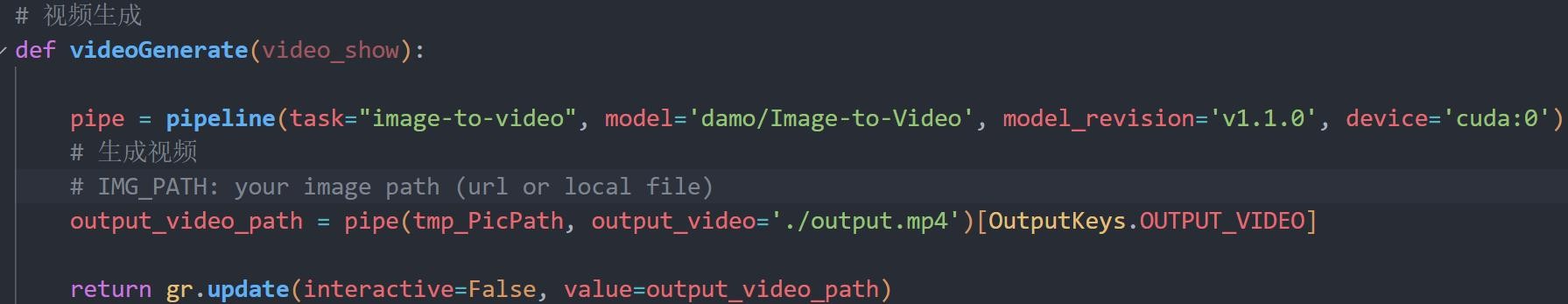
beaAnalysis , an aesthetic evaluation function, calls the Image-Reward model on

the backend to perform aesthetic evaluation on the image and returns the result to the

frontend interface.



qualityAnalysis , a quality detection function, calls the hurt model in the backend to perform quality detection on the image and returns the result to the frontend interface.



videoGenerate , video generation function, calls the xxxx model of the backend, generates the corresponding dynamic video based on the input picture, and returns it to the front-end interface.

### Back-end

#### Functionality

The backend components need to implement the following functions:

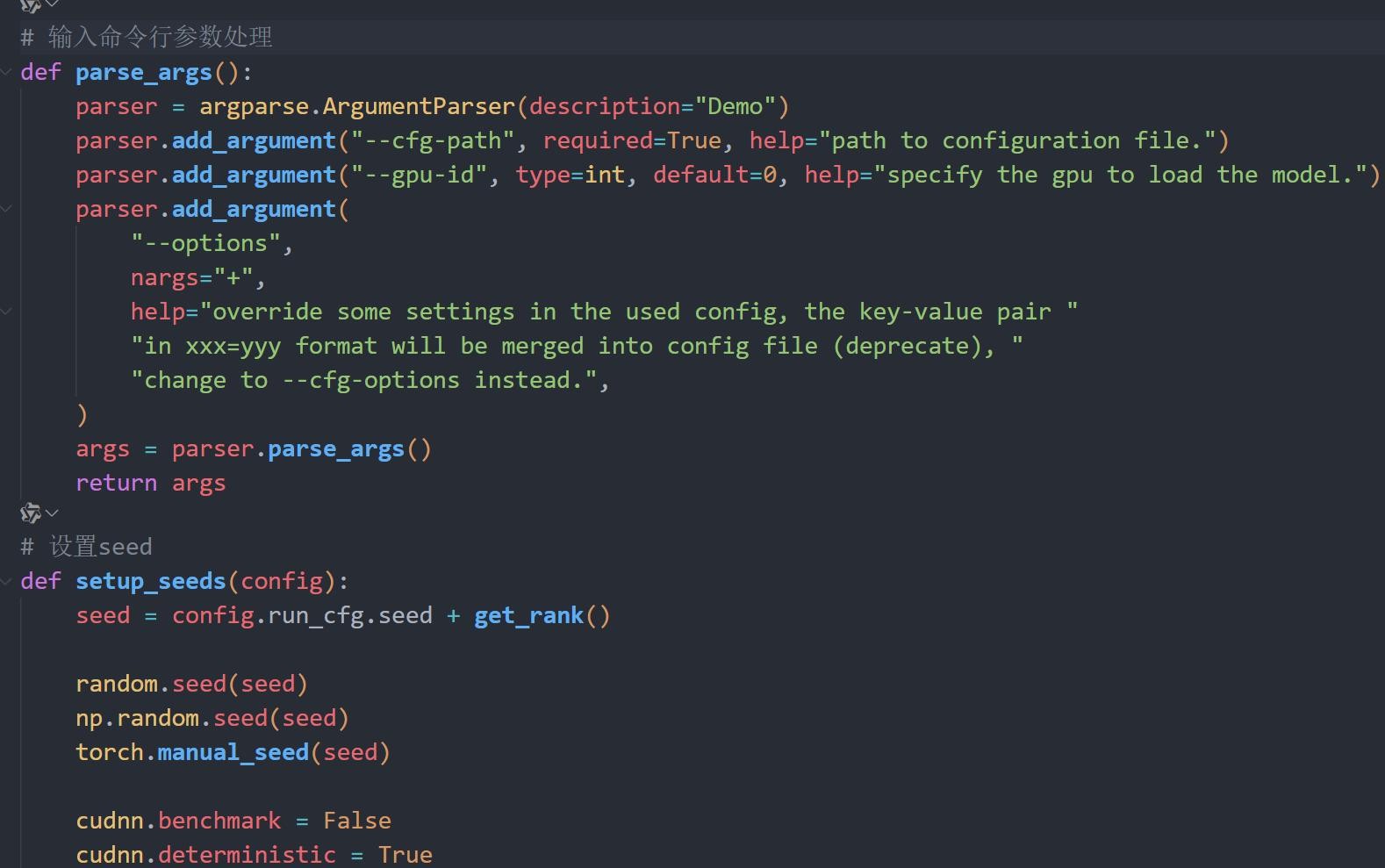
1. Model loading and initialization: When the user enters the front-end interface, the model is loaded and initialized in advance , including MiniGPT , Tongyi Qianwen and other models .
2. Save and call existing results : Save the appreciation results, pictures, etc. in style appreciation , and use these results more practically in subsequent functions .
3. Seamless integration with the front-end: When a user clicks a button on the interface, the information in the front-end is transmitted to the back-end , and the

information returned by the back-end should also be displayed on the front-end . Similarly, when a user clicks the video generation button, the video generation function should be called correctly .

#### Detailed Design

The main backend functions are concentrated in demo.py.

Define the command line parameter processing function parse\_args (), use the argparse library to parse the command line parameters, and return the parsed results. Define the random seed setting function setup\_seeds (config), set the random seed according to the seed and GPU number in the configuration file to ensure the reproducibility of the results.

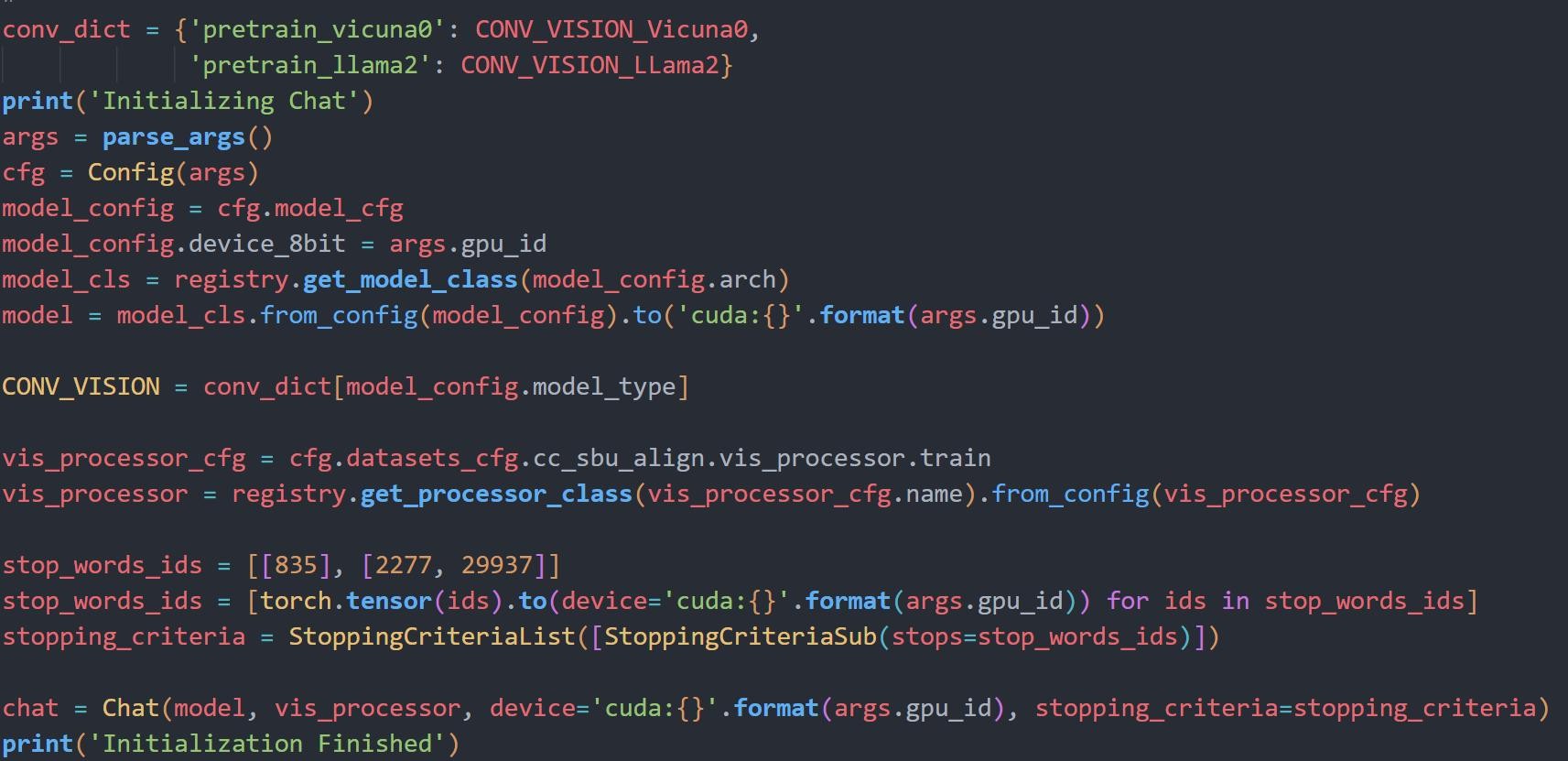


cfg is created according to the command line parameters , and the configuration file is read ; according to the model configuration in the configuration file, the appropriate chat model class model\_cls is selected ; the model object model is created using the model configuration and moved to the specified GPU.

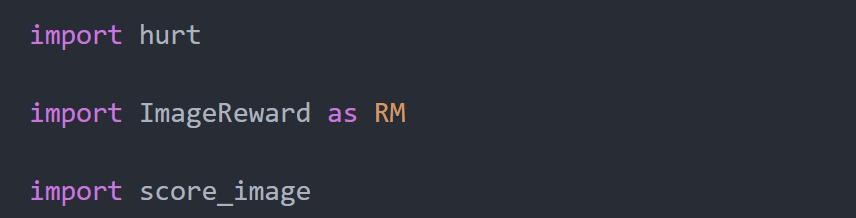
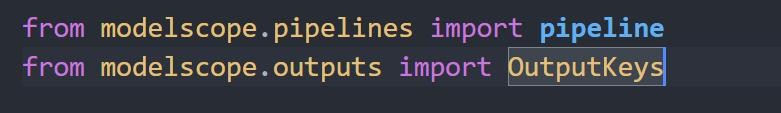
Then initialize other components, select the appropriate chat model type

CONV\_VISION according to the model configuration; create the image processor object vis\_processor according to the configuration file ; set the word ID list of the stop condition ; create the chat object chat, and pass the model, image processor and

stop condition to it.

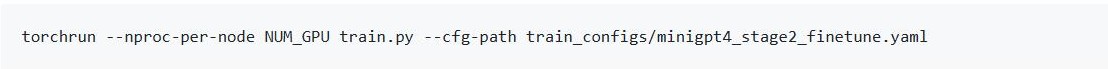


Then initialize the relevant model



### Model Function and Training

MiniGPT-4 training only requires the second step of fine-tuning training. After placing the necessary data sets in the target location, the command line uses the training command to start training:



## Conclusion

This article introduces the project work and related principles of our team, and explains how to use the current artificial intelligence big model to realize the appreciation of dynamic Chinese landscape paintings.

The core models of this paper are the multimodal large model of image description represented by MiniGPT-4 and Image-To-Reward. The two complement each other, one end is used as the input of the whole system, and the other end is used as the output of the whole system. In addition, the output of the image quality detection and image aesthetic evaluation algorithms are used as system parameters, and finally gradio is used as the front-end and back-end framework to integrate the whole system. This project also provides some ideas for the combined use of other large models.

In the future, we will continue to study how to better appreciate paintings, and will use methods such as data set iteration and improved training conditions to obtain better model effects for image description and video generation models.

## References

1. Wang, Peng et al. “OFA: Unifying Architectures, Tasks, and Modalities through a Simple Sequence-to-Sequence Learning Framework”. *International Conference on Machine Learning* (2022).
2. https://blog.csdn.net/JackCrum/article/details/131258219
3. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., & Dong, Y. (2023). ImageReward : Learning and evaluating human preferences for text-to-image generation. *ArXiv , abs/2304.05977* .
4. Lai Wei, Zihao Jiang, Weiran Huang, Lichao Sun .Qing Yuan Research Institute, SEIEE, Shanghai Jiao Tong University ,Lehigh University .InstructionGPT-4: A 200-Instruction Paradigm for Fine-Tuning MiniGPT-4.arXiv:2308.12067v2 [cs.LG] 11 Oct 2023
5. mingyy. (n.d.). Chinese Landscape Paintings dataset. GitHub. <https://github.com/huggingface/datasets/blob/main/>
6. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., & Dong, Y. (2023). ImageReward: Learning and evaluating human preferences for text-to-image generation. arXiv preprint arXiv:2304.05977.
7. Xue, A. (2020). End-to-end Chinese landscape painting creation using generative adversarial networks. arXiv preprint arXiv:2011.05552.
8. Chinese Poetry Foundation. (n.d.). 全 唐 诗 数 据 集 . GitHub.

[https://github.com/chinese-poetry/chinese-poetry/tree/master/%E5%85%A8%E5%94%90%E8%](https://github.com/chinese-poetry/chinese-poetry/tree/master/%E5%85%A8%E5%94%90%E8%AF%97)

AF%97

1. <https://blog.csdn.net/luchentech/article/details/137229498>
2. <https://cloud.tencent.com/developer/article/2378045>
3. [Xiang Wang](https://arxiv.org/search/cs?searchtype=author&query=Wang%2C%2BX), [Hangjie Yuan](https://arxiv.org/search/cs?searchtype=author&query=Yuan%2C%2BH), [Shiwei Zhang](https://arxiv.org/search/cs?searchtype=author&query=Zhang%2C%2BS), [Dayou Chen](https://arxiv.org/search/cs?searchtype=author&query=Chen%2C%2BD), [Jiuniu Wang](https://arxiv.org/search/cs?searchtype=author&query=Wang%2C%2BJ), [Yingya Zhang](https://arxiv.org/search/cs?searchtype=author&query=Zhang%2C%2BY), [Yujun](https://arxiv.org/search/cs?searchtype=author&query=Shen%2C%2BY) Shen, [Deli Zhao](https://arxiv.org/search/cs?searchtype=author&query=Zhao%2C%2BD), [Jingren Zhou](https://arxiv.org/search/cs?searchtype=author&query=Zhou%2C%2BJ).VideoComposer: Compositional Video Synthesis with Motion Controllability.arXiv:2306.02018.