

Image Retrieval System

Prakhar Bhardwaj

Content

1. Understanding the Problem
2. Data Collection
3. Pipeline Structure
4. Generating Image descriptions
5. Image Retrieval System
6. UI Details
7. Tests
8. Challenges
9. Future Improvements

Understanding the Problem

Objective: Develop a Generative AI model for efficiently tagging and categorizing images, enhancing searchability with a focus on user-friendly interface design

Leveraging Existing Technology: Consider the integration of a pre-existing Vision Learning Model (VLM) for its advanced image recognition capabilities along with designing an Image Retrieval system

Data Collection

Diverse Sources for Data Collection:

- Utilized Google web scraping to ensure a wide variety of images
- Leveraged Instagram post datasets for user-generated content
- Accessed digital archive datasets on Kaggle for historical and cultural data

Rationale Behind Data Strategy:

- Decided against generating images via GPT to save time and resources
- Relied instead on publicly available data that already met the project needs

Data Collection - Sample Categories

Adventure Sports,

Analysis of a Pie Chart,

Animals in the Wild,

Athletics,

Car Accident,

Celebrity Instagram Posts,

Cricket Game,

Formula One,

Happy People,

Indian Traditional Apparel,

Maps of Different Regions in the

USA with Analysis,

Maps of Regions in Texas,

People Chatting in Church,

People in a Party,

People in Public Transport,

Indian Traditional Art,

African Traditional Art,

People Protesting,

Sad People,

Street Violence,

Ted Talks,

Tweets on Twitter,

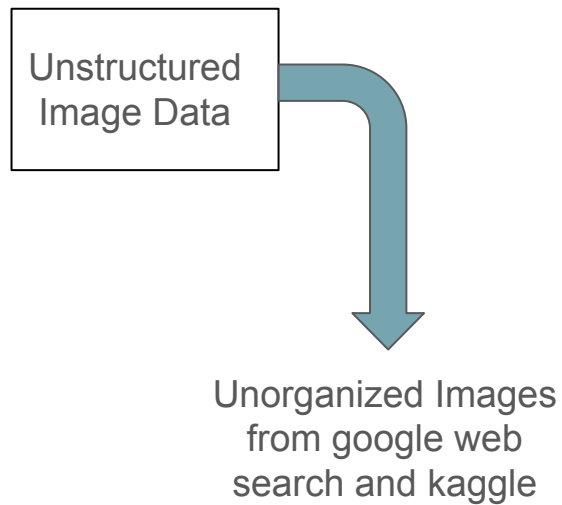
Violent Crowd

Pipeline Structure

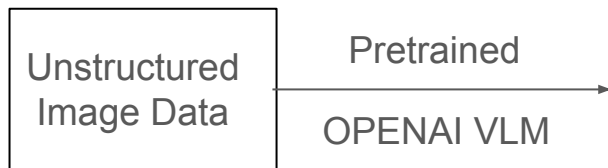


Unstructured
Image Data

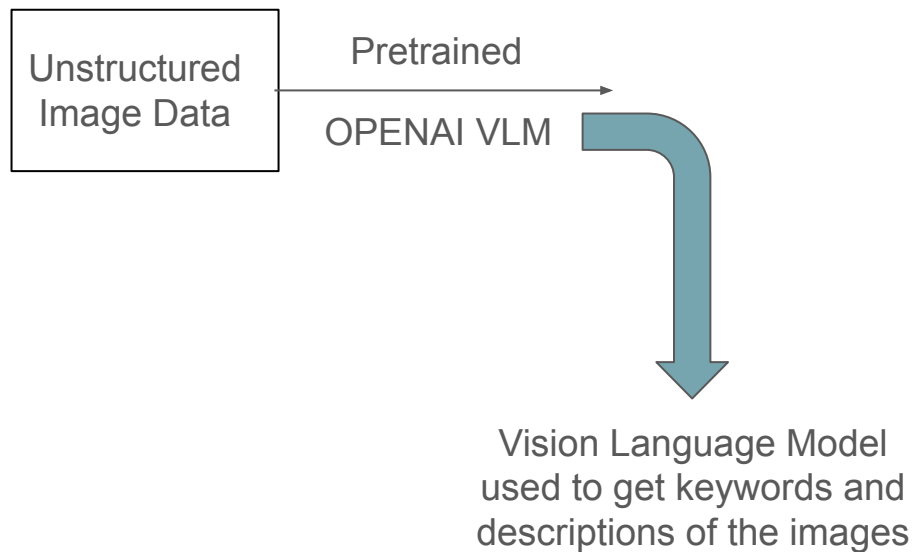
Pipeline Structure



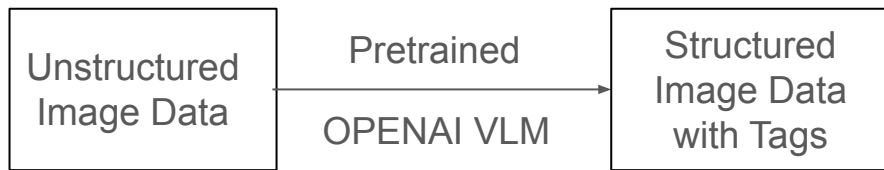
Pipeline Structure



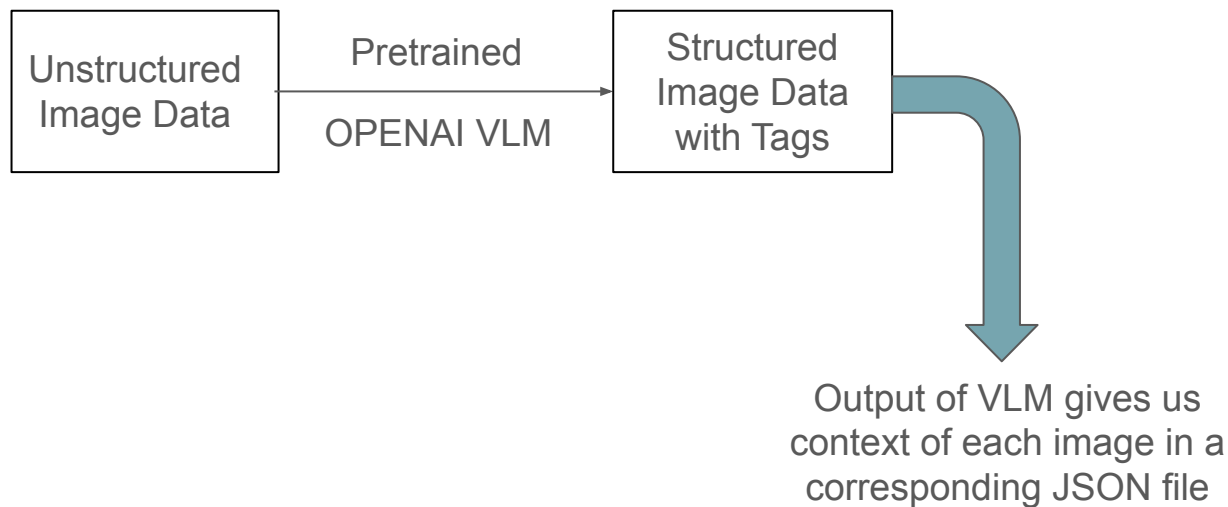
Pipeline Structure



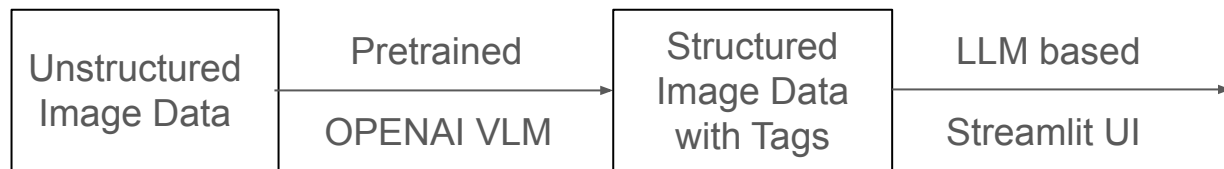
Pipeline Structure



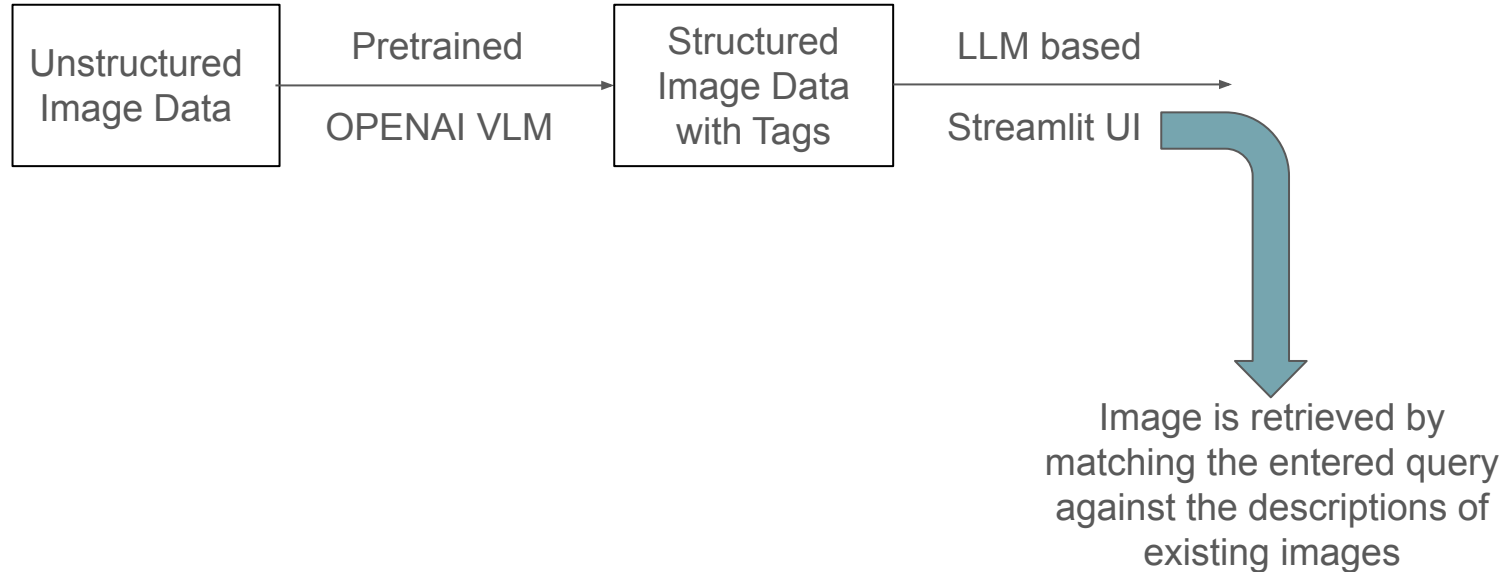
Pipeline Structure



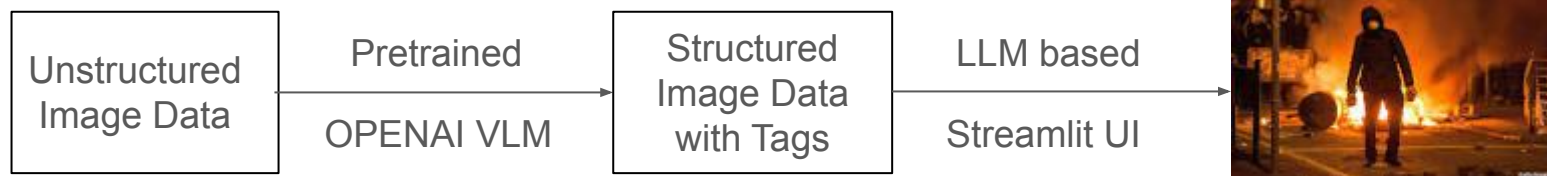
Pipeline Structure



Pipeline Structure



Pipeline Structure



Generating Image Descriptions

1. Leveraged pretrained OPEN AI Vision model through API
2. The code does the following:
 - a. Encoding images from directory into base64 encoding
 - b. Writing a prompt for the model to generate descriptions accordingly
 - c. Generating the JSON file with the same name as the image, having description about the context of the image

Prompts are Important

1. Prompts define how and in what format the model describes the image content
2. Changing the prompt can make the performance of the query matching process really easy
3. Thus, system performance highly depends on the way we frame our prompts
4. We used three different prompts to see how they affect the system performance

Prompts are Important

Prompt #1 -

"Describe the main subjects and objects in this image, including any context or scene details, recognize any text or symbols, and identify emotions or actions depicted. Remember to use only keywords and not the filler words according to the grammar "

Example Guidelines on Prompt by OPENAI

SYSTEM

You will be provided with customer service inquiries that require troubleshooting in a technical support context. Help the user by:

Ask them to check that all cables to/from the router are connected. Note that it is common for cables to come loose over time.

Now you will advise them how to restart their device:

-- If the model number is MTD-327J, advise them to push the red button and hold it for 5 seconds, then wait 5 minutes before testing the connection.

-- If the model number is MTD-327S, advise them to unplug and replug it, then wait 5 minutes before testing the connection.

If the customer's issue persists after restarting the device and waiting 5 minutes, connect them to IT support by outputting {"IT support requested"}.

If the user starts asking questions that are unrelated to this topic then confirm if they would like to end the current chat about troubleshooting and classify their request according to the following scheme:

<insert primary/secondary classification scheme from above here>

Prompts are Important

Prompt #2 -

As an image description generator, your task is to provide concise and accurate descriptions of images within a 20-30 word limit. Follow these guidelines to ensure your descriptions are both informative and succinct:

1. List any people or prominent subjects, noting their positions and actions. If the image is crowded, prioritize the most central or striking figures.
2. Briefly describe significant background elements. Choose details that frame the context of the image effectively.
3. Mention any clear text, symbols, or logos, which are crucial for understanding the image's context or intention.
4. Only comment on the mood or themes like happiness, freedom, success.....etc., if these are directly observable from expressions or actions.
5. Include only what is clearly visible. Do not infer or assume details not explicitly shown in the image.
6. Maintain a neutral tone, avoiding subjective interpretations or emotional language unless directly relevant to the described actions or expressions.

Most importantly your description should balance brevity with informativeness, capturing the essence of the image while adhering to the word limit.

Prompts are Important

Prompt #3 -

As an image tags generator, your task is to provide concise and accurate tags of images. Follow these guidelines to ensure your tags are both informative and succinct:

1. List any people or prominent subjects, noting their positions and actions. If the image is crowded, prioritize the most central or striking figures.
2. Give tags significant background elements. Choose details that frame the context of the image effectively.
3. Give tags for any clear text, symbols, or logos, which are crucial for understanding the image's context or intention.
4. Only give tags on the mood or themes like happiness, freedom, success.....etc., if these are directly observable from expressions or actions.
5. Include only what is clearly visible. Do not infer or assume details not explicitly shown in the image.
6. Maintain a neutral tone, avoiding subjective interpretations or emotional language unless directly relevant to the described actions or expressions.

Most importantly your tags should balance brevity with informativeness, capturing the essence of the image while adhering to the condition that tags must be separated by a space(" ") and for each tag include a few similar words that can be used in place of the original tag

Description Comparison from Different Prompts



Using Prompt #1

"content": "The image shows a group of four young adults, possibly friends, celebrating and having a good time together. They appear to be outdoors at night, as indicated by the dark background which might suggest they are in an urban or street setting. From left to right, there are two women and two men cheering and raising their arms up in the air. The first woman, on the left, is wearing a light-colored top and shorts. Beside her, a man is wearing a dark shirt and light-colored jeans. The third person, another male, is clad in a light-colored, possibly plaid, shirt and dark pants. Lastly, the woman on the far right is wearing a white top and light-colored shorts.\n\nConfetti or petals are captured in mid-air around the group, adding to the festive atmosphere of the photograph and suggesting a celebration, party, or joyful event. There are no discernible text or symbols in the image that can be commented on.\n\nEmotionally, the group looks ecstatic and carefree, possibly dancing or just exuberantly expressing their happiness. Their body language and facial expressions indicate they are having a great time and enjoying the moment."

Description Comparison from Different Prompts



Using Prompt #2

"content": "Group of five young people, celebrating with arms raised amidst falling confetti at night, expressing joy and togetherness."

Description Comparison of Different Prompts



Using Prompt #3

"content": "group celebration party confetti joy excitement night fun friends happiness gathering festivity nightlife cheerful entertainment festive event social togetherness crowd cheer merry jubilation"

Image Retrieval System

1. The system uses an LLM model to understand and compare the descriptions of the images and the user query
2. The code reads descriptions of images from JSON format
3. Each description and the query is turned into an embedding
4. We then compare the query embedding with all the description embeddings to find the closest match
5. After finding the matches, we display them on the UI created with Streamlit
6. If no matches are found, we inform the user that no images could be matched with the query

Image Retrieval System

Initial Model Details -

Embedding model used - 'all-MiniLM-L6-v2', sentence transformer-based model

Similarity measurement - Cosine Similarity

Image Retrieval System

Second Model Details -

Embedding model used - "text-embedding-3-large", an OPEN AI LLM model

Similarity measurement - Cosine Similarity based on the threshold so that only matches greater than the set threshold are retrieved

Image Retrieval System

Third Model Details -

Embedding model used - "text-embedding-3-large", an OPEN AI LLM model

Similarity measurement - FAISS L2 norm for similarity search based on the threshold (calculated using the probability density function of distance)

UI Details

1. Users can input text queries to search for images
2. Users have the option to choose the number of search results they wish to receive: 1, 5, or 10
3. UI filters and displays only relevant images that match the user's search criteria, regardless of the number of results requested
4. Search results are presented in a grid layout with 4 columns
5. Users can download images directly from the search results

Tests

Comparison results of model 1,2 and 3

Enter your search query:
violence scenes

Number of results to retrieve:
5

Search

Result 1 for: "violence scenes"

Result 2 for: "violence scenes"

Result 3 for: "violence scenes"

Result 4 for: "violence scenes"

Result 1 for: "violence scenes"

O/P Model 1

Enter your search query:
violence scene

Number of results to retrieve:
10

Search

Only 8 images found with a similarity score above the threshold.

Result 1 for: "violence scene"

Result 2 for: "violence scene"

Result 3 for: "violence scene"

Result 4 for: "violence scene"

Result 1 for: "violence scene"

Result 2 for: "violence scene"

Result 3 for: "violence scene"

Result 4 for: "violence scene"

O/P Model 2

Enter your search query:
violence scene

Number of results to retrieve:
10

Search

Only 4 images found with a similarity score above the threshold.

Result 1 for: "violence scene"

Result 2 for: "violence scene"

Result 3 for: "violence scene"

O/P Model 3

Tests

Comparison results of model 1,2 and 3

Enter your search query:

various animals with labels

Number of results to retrieve:

5

Search

Result 1 for: "various animals with labels"

Result 2 for: "various animals with labels"

Result 3 for: "various animals with labels"

Result 4 for: "various animals with labels"

Result 1 for: "various animals with labels"

O/P Model 1

Enter your search query:

various animals with labels

Number of results to retrieve:

10

Search

Only 3 images found with a similarity score above the threshold.

Result 1 for: "various animals with labels"

Result 2 for: "various animals with labels"

Result 3 for: "various animals with labels"

O/P Model 2

Enter your search query:

various animals with labels

Number of results to retrieve:

10

Search

Only 2 images found with a similarity score above the threshold.

Result 1 for: "various animals with labels"

Result 2 for: "various animals with labels"

O/P Model 3


Tests


Comparison results of model 1,2 and 3


Enter your search query:
clouds


Number of results to retrieve:
5


Search

Result 1 for: "clouds"


Result 2 for: "clouds"


Result 3 for: "clouds"


Result 4 for: "clouds"


Result 1 for: "clouds"



O/P Model 1


Enter your search query:
clouds


Number of results to retrieve:
10


Search


Only 6 images found with a similarity score above the threshold.


Result 1 for: "clouds"


Result 2 for: "clouds"


Result 3 for: "clouds"


Result 4 for: "clouds"


Result 1 for: "clouds"


Result 2 for: "clouds"



O/P Model 2


Enter your search query:
clouds


Number of results to retrieve:
10

Search

Only 3 images found with a similarity score above the threshold.

Result 1 for: "clouds"


Result 2 for: "clouds"


Result 3 for: "clouds"


O/P Model 3

Tests

Comparison results of model 1,2 and 3

Enter your search query:
a car in an accident

Number of results to retrieve:
5

Search

Result 1 for: "a car in an accident"

Result 2 for: "a car in an accident"

Result 3 for: "a car in an accident"

Result 4 for: "a car in an accident"

O/P Model 1

Enter your search query:
a car in an accident

Number of results to retrieve:
10

Search

Only 6 images found with a similarity score above the threshold.

Result 1 for: "a car in an accident"

Result 2 for: "a car in an accident"

Result 3 for: "a car in an accident"

Result 4 for: "a car in an accident"

O/P Model 2

Enter your search query:
a car in an accident

Number of results to retrieve:
10

Search

Only 4 images found with a similarity score above the threshold.

Result 1 for: "a car in an accident"

Result 2 for: "a car in an accident"

Result 3 for: "a car in an accident"

O/P Model 3

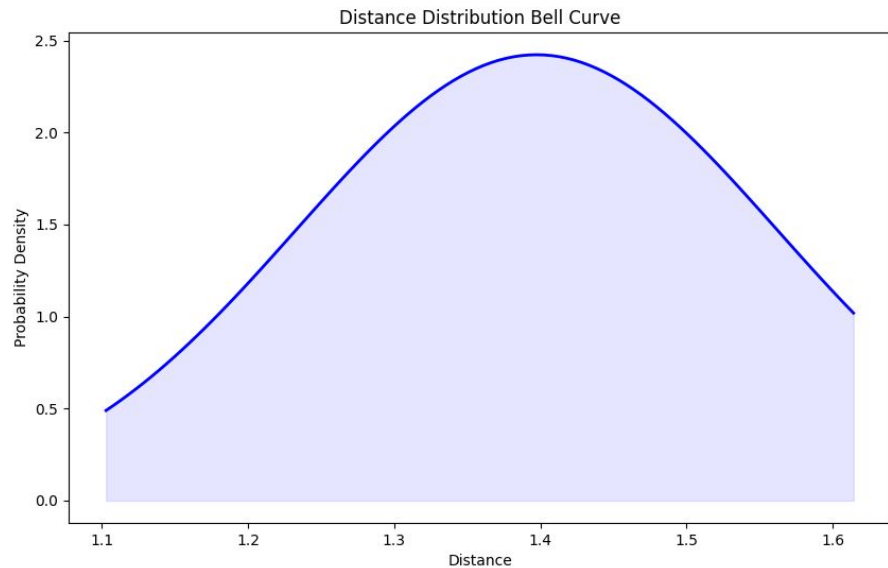
Tests - Best results

Descriptions - Elaborative and comprehensive

Embedding model - OPENAI “text-embedding-3-large”

Similarity Search - FAISS

Threshold - $(\text{mean} - 0.88 \times \text{std})$, based on bell curve



Tests - Best results

Enter your search query:

accident

Number of results to retrieve:

10

Search

Only 1 images found with a similarity score above the threshold.



Result 1 for: "accident"

Enter your search query:

a car in an accident

Number of results to retrieve:

10

Search

Only 1 images found with a similarity score above the threshold.



Result 1 for: "a car in an
accident"

Tests - Best results

Enter your search query:

street violence scene

Number of results to retrieve:

10

Search

Only 1 images found with a similarity score above the threshold.



Result 1 for: "street
violence scene"

Enter your search query:

violence

Number of results to retrieve:

10

Search

Only 2 images found with a similarity score above the threshold.



Result 1 for: "violence"



Result 2 for: "violence"

Enter your search query:

violence scene

Number of results to retrieve:

10

Search

Only 2 images found with a similarity score above the threshold.



Result 1 for: "violence
scene"



Result 2 for: "violence
scene"

Tests - Best results

Enter your search query:

clouds

Number of results to retrieve:

10

Search

Only 2 images found with a similarity score above the threshold.



Result 1 for: "clouds"



Result 2 for: "clouds"

Enter your search query:

various animals with labels

Number of results to retrieve:

10

Search

Only 1 images found with a similarity score above the threshold.



Result 1 for: "various animals with labels"

Challenges

1. Adjusting the threshold for cosine similarity or FAISS similarity search
2. Quantification of the test results using evaluation matrices
3. Making the system robust for huge amount of dataset for varied images
4. The dilemma of spending time in fine tuning the model or writing a better prompt
5. At times we may need to clean the descriptions generated by the VLM

Future Improvements

1. Can have a user directed feedback system where users can indicate the relevance of retrieved images. This data can be used to fine-tune the model or adjust the retrieval algorithm
2. Exploring better techniques to adjust threshold
3. Finding the prompt and description that -
 - a. Contains relevant information about the image
 - b. Does not assume information itself
 - c. Makes threshold adjustment easy
4. Can add multiple language support for the user query and embedding matching
5. Fine tuning the VLM model to generate descriptions according to our needs

Thank You