

## A PROMPT

### A.1 PROMPT FOR INSTRUCTION DATA GENERATION

The prompts we use to generate three types of instruction data are shown in table 5, table 6 and table 7 respectively. Regarding the detailed description type and complex reasoning type data, we will give a manually written example and let ChatGPT generate it using few-shot learning.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Prompt for Detail Description</b></p> <p><b>System:</b> You are an intelligent assistant that can understand video information through text descriptions. You can understand the overall content of the video from the title of the video, the caption of the video. Please describe the video you saw through the information given above. Don't mention the title in your description. Don't copy the original caption. Do not separately describe which objects are included in the video. It is necessary to integrate object information into your description through adjectives or attributive clauses. This description should be between 150 and 200 words.</p> <p><b>User:</b><br/>[title] Guy Scratches Head After Landing Perfect Bowling Strike [Caption] This guy scratched his head in confusion after making a mind-blowing attempt at bowling. He swung his hand to release the ball but accidentally tossed it towards the gutter. However, it spun and turned at the side edges of the lane and then struck all pins in one go.</p> <p><b>ChatGPT:</b> In the video, we see a man wearing a maroon shirt and shorts standing in a bowling alley, holding a bowling ball. First, he swings his hand to release the ball but accidentally tosses it towards the gutter. Next, the ball spins and turns at the side edges of the lane, seemingly heading towards the gutter, but suddenly changes direction and heads towards the pins. . . .</p> <p><b>User:</b><br/>[title] . . . [Caption] . . .</p> |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Table 5: The Prompt for generating instruction data of detail description with few shot learning.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Prompt for Conversation</b></p> <p><b>System:</b> The task is to generate a conversation between two people. One person is watching at a video, and the other person is asking questions about the video. What they see will be provided below with some sentences. Include at least one complex question that requires reasoning and thinking. Only include the questions with certain answers that one can answer with the provided sentences. Make the QA sound like they are seeing the video. Do not use any words that may sound like looking at text instead of images, like "specify", "mention", "description", "text", "provided information", "sentence", "caption", etc. Use words like "see", "look", "view", "show", etc. Format each QA pair in a single line as a JSON dictionary. Do not include any other explanation.</p> <p><b>User:</b><br/>[title] . . . [Caption] . . .</p> |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Table 6: The Prompt for generating instruction data of conversation.

#### A.1.1 PROMPT FOR EVALUATING QA TASK

The prompt that we used ChatGPT to assist in the evaluation in the experiment is shown in the table 8. This prompt will allow the AI to judge the accuracy of the model prediction results and score the consistency of the answer.

## B MORE CASES

### B.1 GENERAL CASES

Figure 8 shows some examples of valley's capabilities in causal reasoning and object identification.

**Prompt for Detail Description**

**System:** You are an AI visual assistant that can analyze a single video. You receive a title of this video and a caption of this video, each describing the same video you are observing. The task is to use the provided title and caption, create a plausible question about the video, and provide the answer in detail. Create complex questions beyond describing the scene. To answer such questions, one should require first understanding the visual content, then based on background knowledge or reasoning, either explain why things are happening that way, or provide guides and help to user's request. Make the question challenging by not including the visual content details in the question so that the user needs to reason about that first. When using the information from the caption, directly explain the scene, and do not mention that the information source is the caption. Always answer as if you are directly looking at the video.

**User:**

[title] Woman Pranks Sister by Covering Inside of Her Whole House in Aluminium Foil  
 [Caption] This woman had gone on a vacation. However, she was shocked when she entered her house on returning. Her sister had covered her whole house with aluminum foil from inside to prank her. She laughed uncontrollably as she saw everything covered in the foil.

**ChatGPT:**

```
{
"question": "Given the sister's initial reaction of uncontrollable laughter upon discovering the
prank, how might this prank affect their relationship in the long run, considering psychological
and social aspects?",
"answer": " From a psychological perspective, humor plays a significant role in maintaining
healthy relationships. The sister's reaction of laughter suggests that she found the prank
amusing, which could enhance their bond. Shared laughter can increase feelings of intimacy
and social cohesion, indicating that the prank may have strengthened their relationship. . . . "
```

**User:**

[title] . . . [Caption] . . .

Table 7: The Prompt for generating instruction data of complex reasoning with few-shot learning.

## B.2 MEME-CAP CASES

Figure 9 shows 2 cases about valley's understanding of picture metaphors.

### Prompt for Detail Description

**System:** You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs.

Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:

---  
## INSTRUCTIONS:

- Focus on the meaningful match between the predicted answer and the correct answer.
- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.

**User:** Please evaluate the following video-based question-answer pair:

Question: {question}

Correct Answer: {answer}

Predicted Answer: {pred}

Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match.

Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where value of 'pred' is a string of 'yes' or 'no' and value of 'score' is in INTEGER, not STRING.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string.

For example, your response should look like this: {'pred': 'yes', 'score': 4.8}."rom inside to prank her. She laughed uncontrollably as she saw everything covered in the foil.

Table 8: The Prompt for QA task evaluation inherited from Video-ChatGPT(Muhammad Maaz & Khan, 2023).



What do you think of this dog?

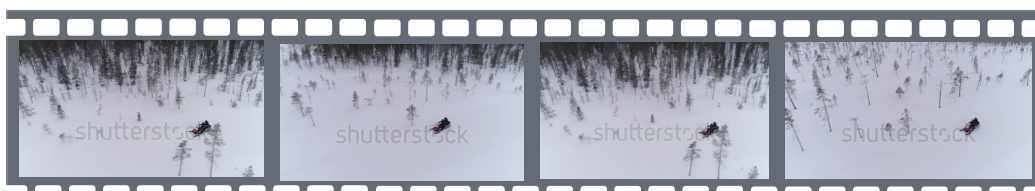
The dog seems to be a fun and intelligent companion. It is able to wear clothes and perform tasks on the computer, which indicates that it has been well-trained and is capable of learning complex tasks. The fact that the dog and the person are engaging in a computer-related activity together shows that they have a strong bond and enjoy spending time together.

Please guess the Inner thoughts of this person.

Based on the video, the person appears to be amazed and entertained by the dog's ability to use the computer. They likely find it amusing and impressive that their pet can not only wear clothes but also participate in activities that were once reserved only for humans. The person may be wondering how the dog was trained to perform these tasks and may be considering ways to further involve the dog in their daily life or creative projects.

Write a short story with this video.

The video captures a lighthearted moment between a person and their dog, showcasing their unique bond and shared interest in technology. The person, initially skeptical of the dog's abilities, is quickly won over by the dog's charming antics and surprising tech skills. As the video progresses, the person and the dog continue to engage in entertaining and educational activities, exploring the world of technology together. In the end, the person realizes that their relationship with their dog has brought a new dimension to their life, and they are grateful for the joy and companionship that the dog has brought into their world.



视频中的场景在哪?

The video shows a snowy forest with trees in the background. The two people are riding their snowmobiles through this scenic landscape.

视频中有几个人?

The video shows a snowy forest with trees in the background. The two people are riding their snowmobiles through this scenic landscape.

能为这个视频写一个故事吗?

Sure! Here's a story for the video: The video shows a winter wonderland, where two friends decided to go for a thrilling ride on their snowmobiles. They both chose their favorite color - red - for their jackets, making them stand out against the snowy backdrop. As they started their journey, they passed by trees and bushes, making the scenery even more picturesque. They rode deeper into the forest, and soon the trees became taller and closer together, creating a feeling of adventure. The friends were having a great time, enjoying the crisp winter air and the exhilarating speed of their snowmobiles. Finally, they emerged from the forest and rode off into the distance, leaving behind a trail of snow and excitement. This video captures the essence of winter fun and the joy of exploring the great outdoors on a snowmobile.

Figure 8: Capture the main content of the video, do the causal inference, and identify objects and relationships in the video. Chinese prompts are also well understood.



**Correct Answer:**

- Meme poster is trying to get healthy but feels temptation eating unhealthy foods.
- Meme poster wants to get healthy but keeps eating fattening King's Hawaiian rolls.
- The meme poster isn't able get fit because of his food habits.
- Poster knows white bread will derail health goal.

**Valley** 🌄: The meme poster is trying to convey a playful and humorous approach to the idea of getting into shape or working out. It features a cartoon character experiencing two contrasting situations: happily receiving a giant yellow ball (which likely symbolizes an enjoyable fitness activity) in the first frame, and then worryingly getting hugged by a giant bag of dinner rolls in the second frame. The two scenes suggest that the character is struggling with their fitness journey, and the humorous scenario is meant to resonate with the audience in a lighthearted manner.

When someone uses a metal spatula on your non stick pan:



**Correct Answer:**

- The poster gets angry that another college student who shares things with him has damaged his non-stick pan.
- Meme poster is sad when someone uses their non-stick pan wrongly.
- The meme poster feels sad for the people who uses spatula on the non stick pan.

**Valley** 🌄: The meme poster aims to convey a sense of anger, frustration or disgust by using a metal faucet as a visual metaphor for a situation that causes unhappiness or annoyance. The phrase "When someone uses a metal spatula on your non-stick pan" suggests that the problem may be persistent or difficult to resolve. The meme poster is likely expressing its dissatisfaction or disappointment with the situation, using the imagery of a spoiled pan to represent the concept.

Figure 9: Two valley examples in the Meme-cap benchmark. The blue box represents the human-annotated understanding of the image, and the green box represents valley's understanding of the image.