ChatGPT Asks, BLIP-2 Answers: Automatic Questioning Towards Enriched Visual Descriptions

Anonymous authors
Paper under double-blind review

Abstract

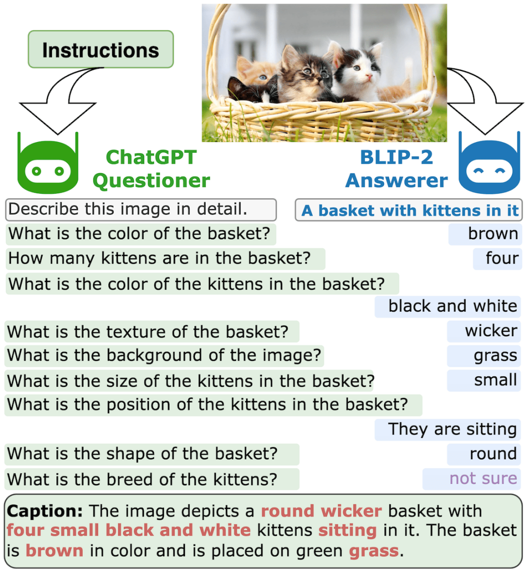
Asking insightful questions is crucial for acquiring knowledge and expanding our understanding of the world. However, the importance of questioning has been largely overlooked in AI research, where models have been primarily developed to answer questions. With the recent advancements of large language models (LLMs) like ChatGPT, we discover their capability to ask high-quality questions when provided with a suitable prompt. This discovery presents a new opportunity to develop an automatic questioning system. In this paper, we introduce ChatCaptioner, a novel automatic-questioning method deployed in image captioning. Here, ChatGPT is prompted to ask a series of informative questions about images to BLIP-2, a strong vision question-answering model. In ChatCaptioner, we investigate whether two AI models, unable to individually describe images in detail, can collaborate through an automated, visually guided dialogue to generate a better and more enriched image description than a single AI model. We conduct human-subject evaluations on common image caption datasets such as COCO, Conceptual Caption, and WikiArt, and compare ChatCaptioner with BLIP-2 as well as ground truth. Our results demonstrate that ChatCaptioner's captions are significantly more informative, receiving three times as many votes from human evaluators as BLLIP-2 alone for providing the most image information. Besides, ChatCaptioner identifies 53% more objects within the image than BLIP-2 alone measured by WordNet synset matching.

1 Introduction

Asking good questions is not only an essential component of effectively acquiring knowledge, but also plays a pivotal role in enhancing our intelligence and expanding our understanding of the world. Taking medical diagnoses as an example, doctors must ask patients targeted questions about their symptoms to gather relevant information and make accurate diagnoses. Likewise, in scientific research, asking insightful questions is paramount to advancing knowledge and discovering new findings that may have far-reaching implications.

However, the primary focus in recent AI research has been on developing models that can better answer questions, like InstructGPT (Ouyang et al., 2022) in Open-Domain Question Answering (Yang et al., 2015; Rajpurkar et al., 2016; Joshi et al., 2017) and BLIP-2 (Li et al., 2023) in Visual Question Answering (Antol et al., 2015; Goyal et al., 2017; Hudson & Manning, 2019). Despite the significant progress in the question-answering models, their effectiveness in providing useful information is heavily reliant on the quality of the questions they receive. In essence, these models depend on humans to ask insightful questions that can direct their generation of informative answers. If we have an automatic questioning machine that keeps asking informative questions, the human questioners can be replaced and the question-answering models can be guided to provide more valuable knowledge automatically.

Recent studies (Wei et al., 2022a; Ouyang et al., 2022; Wei et al., 2020; Kojima et al.) have highlighted the impressive zero-shot learning abilities of Large Language Models (LLMs) that are fine-tuned to follow instructions. These LLMs can perform new tasks in a zero-shot manner when presented with well-crafted instruction prompts. We discover that such LLMs like ChatGPT (OpenAI, 2022) have the ability to keep ask-



ChatGPT Questioner	BLIP-2 Answerer
Describe this image in detail.	**A basket with kittens in it**
What is the color of the basket?	brown
How many kittens are in the basket?	four
What is the color of the kittens in the basket?	black and white
What is the texture of the basket?	wicker
What is the background of the image?	grass
What is the size of the kittens in the basket?	small
What is the position of the kittens in the basket?	They are sitting
What is the shape of the basket?	round
What is the breed of the kittens?	not sure

Caption: The image depicts a **round wicker** basket with **four small black and white** kittens **sitting** in it. The basket is **brown** in color and is placed on green **grass**.

Figure 1: **Example of the dialog between ChatGPT and BLIP-2:** BLIP-2 fails to provide a detailed description in the first message exchange. More details about the image highlighted in red are obtained through multiple conversational interactions between the two models. Eventually, the questioner is able to produce a *more detailed* caption about the image by focusing on multiple aspects of the image.

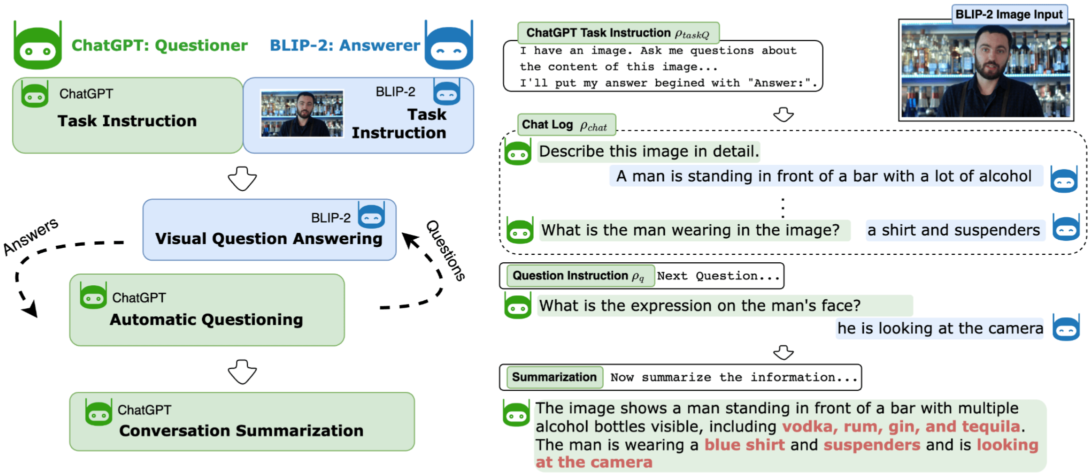
ing new and contextually relevant questions when properly designed prompts are given. With this capability in place, building an effective automatic questioning machine is now a feasible task.

Based on our findings, we design an automatic questioning system on ChatGPT and integrate it into image captioning, where strong vision-language models like BLIP-2 (Li et al., 2023) are available to answer image-related questions. Our method, named ChatCaptioner, generates more informative and enriched image captions by asking relevant questions to incrementally gain more information. In detail, we design a prompting system that encourages ChatGPT to ask a series of informative questions that maximize its knowledge of the image, building on previous questions and answers. Note that ChatGPT is a pure language model and cannot “see” any visual information. We present the inquired image to BLIP-2 and set it as the question answerer. At the end of the conversation, ChatGPT is prompted to summarize the discussion into a few sentences as the final enriched image description. An example of the conversation between ChatGPT and BLIP-2 and the final caption is shown in Fig.1.

In our experiments, we aim to investigate whether ChatCaptioner, an automated visual dialogue system between a large language model ChatGPT and a visual language model BLIP-2, can acquire more visual information compared to the output of the visual language model alone. We evaluate ChatCaptioner's captions on sampled images from COCO (Lin et al., 2014), WikiArt (Saleh & Elgammal, 2015), and CC (Sharma et al., 2018) datasets based on the human subject evaluation experiment. Compared to BLIP-2's state-of-the-art direct captioning performance, ChatCaptioner receives three times as many votes from human evaluators for providing richer image information. Besides, ChatCaptioner identifies 53% more objects than BLIP-2 alone within the image. Results verify the benefit of good questions to acquire more knowledge from existing AI models and the effectiveness of modern LLMs to serve as zero-shot automatic questioners.

2 Related Works

Learning to Ask Questions Question generation (Mostow & Chen, 2009; Heilman & Smith, 2010) is the task of generating a question from a given passage and an answer. Recent methods like Jia et al. (2020); Xiao et al. (2020); Liu et al. (2019); Ghanem et al. (2022); Wang et al. (2022a) have explored different



```
I have an image. Ask me questions about
the content of this image...
I'll put my answer begined with "Answer:".
```

Describe this image in detail.

A man is standing in front of a bar with a lot of alcohol

⋮

What is the man wearing in the image?

a shirt and suspenders

`Next Question...`

What is the expression on the man's face?

he is looking at the camera

`Now summarize the information...`

The image shows a man standing in front of a bar with multiple alcohol bottles visible, including **vodka, rum, gin, and tequila**. The man is wearing a **blue shirt** and **suspenders** and is **looking at the camera**

Figure 2: ChatCaptioner contains a questioner ChatGPT and an answerer BLIP-2. We start with a task instruction to set the context and outline the problem for ChatGPT. Once the answer from BLIP-2 is received, we prompt the chat log and question instruction to ChatGPT to continue asking information-acquiring questions. Finally, we provide a summarization instruction to ChatGPT to conclude the conversation as the image captions. Enriched details are highlighted in red.

neural network architectures and training strategies for better performance. However, in cases when we do not have the answer and need to ask questions for the answers, such methods are not applicable. Visual Question Generation (Mostafazadeh et al., 2016; Zhang et al., 2016) is a task aimed at generating natural and engaging questions for a given image. Several works like Patro et al. (2018; 2020); Li et al. (2018); Jain et al. (2017); Vedd et al. (2021); Shen et al. (2019) have been proposed to solve this task. They focus on generating independent questions only and do not have the ability to keep asking new questions based on the previous questions. Our work differs from previous studies significantly. First, we focus on acquiring more knowledge via the generated questions, instead of just generating them. Secondly, our method can keep asking new and relevant questions based on previous questions. Third, our approach leverages modern large language models and requires zero training for questioning.

Large Language Model and Prompting Recent research (Brown et al., 2020; Kojima et al.; Wei et al., 2020; 2022b;a; Chung et al., 2022; Ouyang et al., 2022) has revealed the abilities of Large Language Models (LLMs) like GPT-3 (Brown et al., 2020) or PaLM (Chowdhery et al., 2022) to solve versatile tasks specified by prompting. For example, GPT-3 (Brown et al., 2020) shows the capability to learn new tasks by providing a few task examples provided in the prompt, named in-context learning. Moreover, Chain-of-Thought methods (Kojima et al.; Wei et al., 2020) demonstrate that explicitly asking LLM to solve tasks step-by-step in the prompt improves the performance significantly. Additionally, FLAN (Wei et al., 2022a; Chung et al., 2022) demonstrates that LLMs with instruction tuning can accomplish new tasks in a zero-shot manner. Further studies, including InstructGPT (Ouyang et al., 2022) and ChatGPT (OpenAI, 2022), show that the performance of LLMs can be enhanced even further by using reinforcement learning from human feedback (Christiano et al., 2017; Stiennon et al., 2020). In our work, we leverage the instruction-following ability of LLMs and design prompts that enable ChatGPT to keep asking new questions about images.

Image Captioning and Visual Question Answering Recent research in vision and language pertaining (Chen et al., 2022; Tsimpoukelli et al., 2021; Alayrac et al., 2022; Wang et al., 2022b; Li et al., 2022; 2023) has advanced the performance for image captioning and visual question answering (VQA) by a large margin. For example, VisualGPT (Chen et al., 2022) shows the benefits of initialization with pretrained language models for more data-efficient training. Frozen (Tsimpoukelli et al., 2021) extend it by finetuning a vision encoder and aligning it with a frozen LLM. BEiT-3 (Wang et al., 2022b) and BLIP (Li et al., 2022) pretrain models using unified transformer architecture. Flamingo (Alayrac et al., 2022) proposes a cross-attention design to align vision and language. BLIP-2 (Li et al., 2023) introduces a lightweight Q-Former that converts visual features into tokens that can be directly understood by a frozen LLM, and achieves impressive results on both image captioning and VQA tasks. In our work, our automatic questioning mechanism leverages the VQA capability of BLIP-2 to extract additional image information and enrich the image captions beyond the original BLIP-2 captions.

3 ChatCaptioner

In ChatCaptioner, we design an automatic questioning mechanism based on ChatGPT's zero-shot instruction-following ability to keep asking informative questions about images. BLIP-2, the vision-language model, then provides new image information according to the asked questions. Finally, ChatGPT is prompted to summarize the chat history and generate the final image captions with rich details. An overview of our method is demonstrated in Fig.2.

3.1 Automatic Questioning

To activate the questioning ability of ChatGPT, we design a prompting system that enables ChatGPT to generate questions based on previous chat logs. Our prompting system for ChatGPT contains three components: a task instruction for explaining the task ρ_{taskQ}, a chat log to store previous questions and answers ρ_{chat}, a question instruction for generating high-quality questions ρ_q. Each question is generated given the context $\rho_{taskQ}+\rho_{chat}+\rho_q$. In addition, we design a question-trimming mechanism for the automatic post-processing of the generated questions.

ChatGPT Task Instruction ρ_{taskQ} The task instruction ρ_{taskQ} sets the context and outlines the task that ChatGPT is required to perform. ρ_{taskQ} directs ChatGPT to generate questions that aim to extract as much information as possible about an image. ρ_{taskQ} is designed as follows:

I have an image. Ask me questions about the content of this image. Carefully asking me informative questions to maximize your information about this image content. Each time ask one question only without giving an answer. Avoid asking yes/no questions. I'll put my answer beginning with "Answer:".

The prompt "Each time ask one question only without giving an answer" in ρ_{taskQ} is designed to instruct ChatGPT to generate only one question per round. Without this prompt, ChatGPT may generate a set of questions at once, rather than generating them one by one in response to each answer.

Chat Log ρ_{chat} After the task instruction ρ_{taskQ}, we include a chat log ρ_{chat} containing all the previous questions and answers. ρ_{chat} follows a template for each Question-Answer round, which is as follows:

Question: $\langle$*question*$\rangle$ *Answer:* $\langle$*answer*$\rangle$

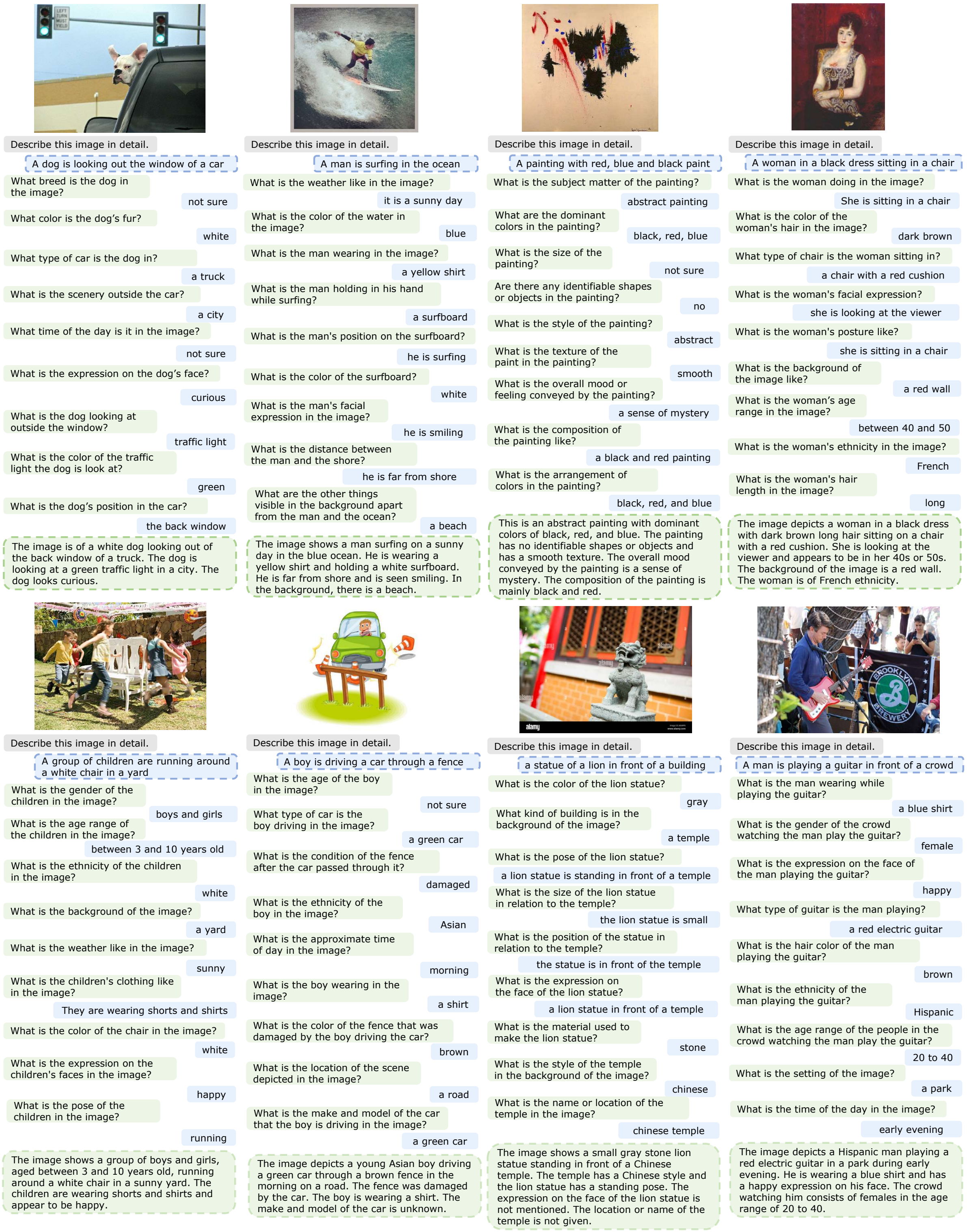
Note that we hard-code the first question as "Describe the image in detail" to start the conversation. Based on the first response of BLIP-2, which provides a brief initial description of the image, ChatGPT is prompted to ask follow-up questions to extract more information about the image.

Question Instruction ρ_q To guide ChatGPT in generating new questions, we provide a question instruction ρ_q before each question generation. ρ_q, located after the chat log, cues ChatGPT to generate a new question and aims to ensure that the questions are of high quality. It's designed as follows:

Next Question. Avoid asking yes/no questions. Question:

The prompt "Next Question" in ρ_q is critical to ensure that ChatGPT continues to ask questions. Without it, ChatGPT may produce undesired sentences after a few Question-Answer rounds. Additionally, we notice that ChatGPT prefers asking yes/no questions which are usually not as informative as other questions. We therefore add the prompt "Avoid asking yes/no questions" to reduce the generation of yes/no questions.

Question Trimming Despite our explicit instruction to not answer the question itself, we observe that sometimes ChatGPT fabricates an answer after asking the question. Fortunately, we find that these fab-



Describe this image in detail.

A dog is looking out the window of a car

What breed is the dog in the image?

not sure

What color is the dog's fur?

white

What type of car is the dog in?

a truck

What is the scenery outside the car?

a city

What time of the day is it in the image?

not sure

What is the expression on the dog's face?

curious

What is the dog looking at outside the window?

traffic light

What is the color of the traffic light the dog is look at?

green

What is the dog's position in the car?

the back window

The image is of a white dog looking out of the back window of a truck. The dog is looking at a green traffic light in a city. The dog looks curious.

Describe this image in detail.

A man is surfing in the ocean

What is the weather like in the image?

it is a sunny day

What is the color of the water in the image?

blue

What is the man wearing in the image?

a yellow shirt

What is the man holding in his hand while surfing?

a surfboard

What is the man's position on the surfboard?

he is surfing

What is the color of the surfboard?

white

What is the man's facial expression in the image?

he is smiling

What is the distance between the man and the shore?

he is far from shore

What are the other things visible in the background apart from the man and the ocean?

a beach

The image shows a man surfing on a sunny day in the blue ocean. He is wearing a yellow shirt and holding a white surfboard. He is far from shore and is seen smiling. In the background, there is a beach.

Describe this image in detail.

A painting with red, blue and black paint

What is the subject matter of the painting?

abstract painting

What are the dominant colors in the painting?

black, red, blue

What is the size of the painting?

not sure

Are there any identifiable shapes or objects in the painting?

no

What is the style of the painting?

abstract

What is the texture of the paint in the painting?

smooth

What is the overall mood or feeling conveyed by the painting?

a sense of mystery

What is the composition of the painting like?

a black and red painting

What is the arrangement of colors in the painting?

black, red, and blue

This is an abstract painting with dominant colors of black, red, and blue. The painting has no identifiable shapes or objects and has a smooth texture. The overall mood conveyed by the painting is a sense of mystery. The composition of the painting is mainly black and red.

Describe this image in detail.

A woman in a black dress sitting in a chair

What is the woman doing in the image?

She is sitting in a chair

What is the color of the woman's hair in the image?

dark brown

What type of chair is the woman sitting in?

a chair with a red cushion

What is the woman's facial expression?

she is looking at the viewer

What is the woman's posture like?

she is sitting in a chair

What is the background of the image like?

a red wall

What is the woman's age range in the image?

between 40 and 50

What is the woman's ethnicity in the image?

French

What is the woman's hair length in the image?

long

The image depicts a woman in a black dress with dark brown long hair sitting on a chair with a red cushion. She is looking at the viewer and appears to be in her 40s or 50s. The background of the image is a red wall. The woman is of French ethnicity.

Describe this image in detail.

A group of children are running around a white chair in a yard

What is the gender of the children in the image?

boys and girls

What is the age range of the children in the image?

between 3 and 10 years old

What is the ethnicity of the children in the image?

white

What is the background of the image?

a yard

What is the weather like in the image?

sunny

What is the children's clothing like in the image?

They are wearing shorts and shirts

What is the color of the chair in the image?

white

What is the expression on the children's faces in the image?

happy

What is the pose of the children in the image?

running

The image shows a group of boys and girls, aged between 3 and 10 years old, running around a white chair in a sunny yard. The children are wearing shorts and shirts and appear to be happy.

Describe this image in detail.

A boy is driving a car through a fence

What is the age of the boy in the image?

not sure

What type of car is the boy driving in the image?

a green car

What is the condition of the fence after the car passed through it?

damaged

What is the ethnicity of the boy in the image?

Asian

What is the approximate time of day in the image?

morning

What is the boy wearing in the image?

a shirt

What is the color of the fence that was damaged by the boy driving the car?

brown

What is the location of the scene depicted in the image?

a road

What is the make and model of the car that the boy is driving in the image?

a green car

The image depicts a young Asian boy driving a green car through a brown fence in the morning on a road. The fence was damaged by the car. The boy is wearing a shirt. The make and model of the car is unknown.

Describe this image in detail.

a statue of a lion in front of a building

What is the color of the lion statue?

gray

What kind of building is in the background of the image?

a temple

What is the pose of the lion statue?

a lion statue is standing in front of a temple

What is the size of the lion statue in relation to the temple?

the lion statue is small

What is the position of the statue in relation to the temple?

the statue is in front of the temple

What is the expression on the face of the lion statue?

a lion statue in front of a temple

What is the material used to make the lion statue?

stone

What is the style of the temple in the background of the image?

chinese

What is the name or location of the temple in the image?

chinese temple

The image shows a small gray stone lion statue standing in front of a Chinese temple. The temple has a Chinese style and the lion statue has a standing pose. The expression on the face of the lion statue is not mentioned. The location or name of the temple is not given.

Describe this image in detail.

A man is playing a guitar in front of a crowd

What is the man wearing while playing the guitar?

a blue shirt

What is the gender of the crowd watching the man play the guitar?

female

What is the expression on the face of the man playing the guitar?

happy

What type of guitar is the man playing?

a red electric guitar

What is the hair color of the man playing the guitar?

brown

What is the ethnicity of the man playing the guitar?

Hispanic

What is the age range of the people in the crowd watching the man play the guitar?

20 to 40

What is the setting of the image?

a park

What is the time of the day in the image?

early evening

The image depicts a Hispanic man playing a red electric guitar in a park during early evening. He is wearing a blue shirt and has a happy expression on his face. The crowd watching him consists of females in the age range of 20 to 40.

Figure 3: Qualitative examples of the chat log and the captions from ChatCaptioner in various images from COCO (Lin et al., 2014), WikiArt (Saleh & Elgammal, 2015), and CC (Sharma et al., 2018). Compared to the initial description from BLIP-2, questions from ChatGPT extract additional image information and lead to enriched final captions.

ricated answers always begin with the text "Answer:", following the template specified in the prompt. Therefore, we automatically remove these fabricated answers by discarding the generated text starting from "Answer:".

3.2 Question Answering

Similar to ChatGPT, our BLIP-2 prompting mechanism consists of three components: a task instruction ρ_{taskA}, the chat log ρ_{chat} same as the ChatGPT one, and an answer instruction ρ_a. Each answer generation is prompted by $\rho_{taskA} + \rho_{chat} + \rho_a$. Also, we have an answer-trimming mechanism for post-processing.

BLIP-2 Task Instruction ρ_{taskA} We design the BLIP-2 task instruction ρ_{taskA} to alleviate the issue of hallucinating non-existent information in the image. ρ_{taskA} includes an uncertainty prompt "If you are not sure about the answer, say you don't know honestly" that encourages BLIP-2's honest admission of lack of knowledge. The instruction is as follows:

Answer given questions. If you are not sure about the answer, say you don't know honestly. Don't imagine any contents that are not in the image.

Answer Instruction ρ_a After the chat log ρ_{chat}, we provide a straightforward answer instruction to guide BLIP-2's answering process. The instruction is structured as follows:

Question: $\langle$*question*$\rangle$ *Answer:*

Answer Trimming Similar to ChatGPT, BLIP-2 occasionally generates a question after providing an answer. As the LLM backend of BLIP-2, the FLAN-T5 model (Chung et al., 2022), has a much weaker questioning ability than ChatGPT shown later in the question analysis in the experiment section, we automatically filter out these questions by discarding any texts starting with "Question:".

3.3 Context Summarizing

To obtain a concise summary of the conversation between ChatGPT and BLIP-2 as the final image caption, we use a summarization instruction after the conversation. This instruction, located after the chat log, prompts ChatGPT to generate a summary using the following structure:

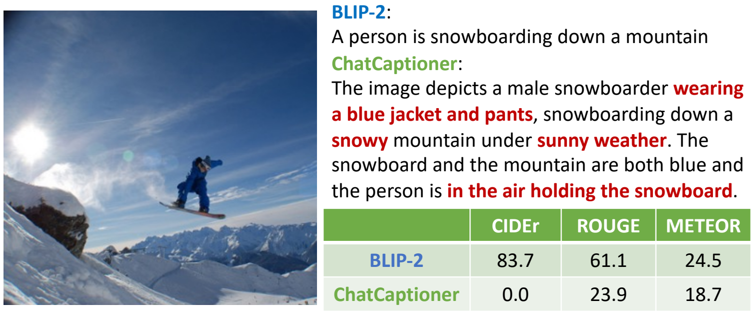
Now summarize the information you get in a few sentences. Ignore the questions with answers no or not sure. Don't add information. Don't miss information. Summary:

4 Experiments

We explore ChatCaptioner from various perspectives through a series of experiments, including an assessment of the informativeness and accuracy of its captions, as well as an analysis of the quality of its generated questions. Qualitative results including chat logs and final captions on various images from different datasets are shown in Fig.3.

Details of Model Deployment. For our experiments, we use the ChatGPT model *"gpt-3.5-turbo"* available on the OpenAI API (cha, 2023). This model is the most powerful GPT-3.5 model accessible through the API during our project. For BLIP-2, we use the biggest version containing a FLAN-T5 (Chung et al., 2022) language model with 11 billion parameters and a ViT-G/14 model from EVA-CLIP (Fang et al., 2022). In all experiments, BLIP-2 answers 10 questions per image, with the first question being hard-coded as *"Describe the image in detail."*. The remaining 9 questions are from ChatGPT, unless otherwise specified.

Limitation of Traditional Metrics. The conventional image captioning metrics such as(Vedantam et al., 2015), ROUGE (Lin & Hovy, 2002), and METEOR (Banerjee & Lavie, 2005) are frequently employed to measure the quality of captions. However, the usage of these metrics to evaluate ChatCaptioner can be limiting, because these metrics are designed to measure the similarity between the tested caption and reference captions, assuming that the reference captions are perfect image descriptions. Since ChatCaptioner



BLIP-2:
A person is snowboarding down a mountain
ChatCaptioner:
The image depicts a male snowboarder **wearing a blue jacket and pants**, snowboarding down a **snowy** mountain under **sunny weather**. The snowboard and the mountain are both blue and the person is **in the air holding the snowboard**.

	CIDEr	ROUGE	METEOR
BLIP-2	83.7	61.1	24.5
ChatCaptioner	0.0	23.9	18.7

Figure 4: An example of the limitation of traditional image caption metrics. Although ChatCaptioner extracts more image details highlighted in red compared to BLIP-2, it receives much lower scores under traditional metrics.

Table 1: Human votes on the captions containing the most image information.

Methods	COCO	WikiArt	CC	OI-LN	Avg.
GT	26%	14%	8.5%	33.5%	20.5%
BLIP-2	21%	12.5%	23%	6.5%	15.8%
Ours	**53%**	**73.5%**	**68.5%**	**63.8%**	**65%**

generates captions that contain a greater level of detail than the reference captions, the metrics may yield low similarity scores, leading to inferior overall performance scores compared to other image captioning approaches like BLIP-2. This issue is depicted in Fig.4. Thus, in our experiments, we primarily rely on human assessments to comprehensively analyze the performance of ChatCaptioner from various perspectives. The design of all the human evaluation interfaces is presented in the supplementary.

4.1 Information Analysis

Does ChatCaptioner extract more information from the image? We design an experiment to evaluate whether ChatCaptioner is able to generate captions with more information about the image than BLIP-2 alone and ground truth captions. We randomly selected 100 photos from the COCO (Lin et al., 2014) validation set, 100 artworks from WikiArt (Saleh & Elgammal, 2015) dataset with ground truth captions from ArtEmis (Achlioptas et al., 2021), 100 internet images from the Conceptual Captions (CC) (Sharma et al., 2018) validation dataset, and 100 images with detailed and long ground truth captions from the Open Image Localized Narratives (OI-LN) (Pont-Tuset et al., 2020) dataset. Human evaluators on Amazon Mechanical Turk are presented with an image and four captions - one from our method, one from BLIP-2, one ground truth caption, and one fake caption for quality control. Evaluators are asked to pick the caption that offers the richest information about the image. Results are demonstrated in Tab.1. On average, ChatCaptioner receives three to four times as many votes as pure BLIP-2's captions and ground truth captions, showing that by combining two AI models via questioning and asking, the system can provide more detailed image description than a single AI model alone.

How many objects in images can ChatCaptioner discover? We randomly sampled 200 images from Pascal VOC (Everingham et al., 2010) and considered all class labels in the segmentation masks as the ground truth objects. We then assessed how many of these objects are included in the captions. We utilize WordNet from NLTK (Bird et al., 2009) to find words with similar semantic meanings based on the Wu-Palmer Similarity of their synsets. Tab.2 presents the experimental results, where 1154 objects are identified in the 200 sampled images. BLIP-2 covers only 383 of them, while ChatCaptioner increases the coverage by 53% to 586, suggesting that the automatic questioning helps BLIP-2 find more objects in the images.

4.2 Correctness Analysis

How accurate are the captions from ChatCaptioner? We conducted a human evaluation where evaluators were presented with an image and a generated caption, as well as all Q&A between ChatGPT

Table 2: Numbers of objects discovered by captions.

Methods	**Covered/All**	**Ratio**	**Improved**
BLIP-2	383/1154	33.2%	-
Ours	**586**/1154	**50.8%**	**53.0%**

Table 3: Correctness Analysis. BLIP-2 correctly answers 66.7% of ChatGPT's questions. 81% of the final captions are deemed correct by humans. Besides, 94% of the wrong captions are caused by BLIP-2's wrong answers.

	COCO	**WikiArt**	**CC**	**Avg.**
Answer Correct Rate	64%	73%	63%	66.7%
Caption Correct Rate	77%	78%	88%	81%
Issues From BLIP-2	100%	82%	100%	94%

and BLIP-2. The evaluators need to verify the correctness of the caption, select incorrect answers from BLIP-2, and judge whether the incorrectness can be attributed to the wrong answers. The experiments were performed on samples from COCO (Lin et al., 2014), WikiArt (Saleh & Elgammal, 2015), and CC (Sharma et al., 2018) datasets. Each image was evaluated by 4 different evaluators. Results presented in Tab.3 reveal that about 80% of the captions are deemed correct. Moreover, BLIP-2 answers around 67% of the questions correctly. Among the incorrect captions, 94% are caused by BLIP-2's wrong answers, suggesting that BLIP-2 is the primary source of incorrectness. This implies that using a more powerful VQA model may help to enhance the overall performance of the system in the future.

Does BLIP-2 know it doesn't know? BLIP-2 usually makes up answers if the question cannot be answered based on the image. In other words, BLIP-2 doesn't know that it doesn't know this information. To mitigate this issue, we incorporate an uncertainty prompt *"If you are not sure about the answer, say you don't know honestly."* in our BLIP-2 task instruction ρ_{taskA}. Two examples showing the effectiveness of the uncertainty prompt are demonstrated in Fig.5 and more examples are in Fig.3 and the supplementary.

How effective is the uncertainty prompt? We randomly selected 200 images from the CC (Sharma et al., 2018) and collected 1,800 questions based on these images. We then identify 147 questions that BLIP-2 is uncertain about, present these questions to human evaluators, and ask them to answer based on the image content. Results in Tab.4 demonstrate that approximately 60% of these questions are deemed unanswerable based on the image. For the remaining answerable questions, BLIP-2 cannot correctly answer 30 of them. In total, without the uncertainty prompt, BLIP-2 will generate 118 incorrect answers out of 147 uncertain questions, resulting in an error rate of approximately 80%. In addition, out of the original 1800 questions, BLIP-2 has 674 wrong answers. Taking the 147 potential wrong answers avoided by the uncertainty prompt into account, the uncertainty prompt reduces about 15% of the wrong answers.

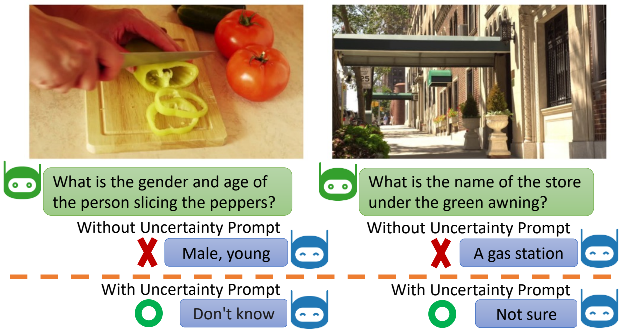

What is the gender and age of the person slicing the peppers?
Without Uncertainty Prompt
Male, young
With Uncertainty Prompt
Don't know
What is the name of the store under the green awning?
Without Uncertainty Prompt
A gas station
With Uncertainty Prompt
Not sure

Figure 5: Examples of BLIP-2's answers with and without the uncertainty prompt. The uncertainty prompt helps BLIP-2 avoid making up an answer when it encounters questions that cannot be answered based on the image.

Table 4: Analysis on questions that BLIP-2 is unsure about. 60% deemed unanswerable by humans. 20% cannot be correctly answered by BLIP-2. Overall, BLIP-2 makes mistakes on 80% of these questions.

	Total Uncertain Questions	Unanswerable Questions	Answerable But Wrong	Avoided Bad Answers
Num.	147	88	30	118
Ratio	-	60%	20%	80%

Table 5: Number of unique questions. InstructGPT and ChatGPT excel at generating diverse questions and rarely repeating questions within a dialogue.

Unique Q/Total Q	OPT 6.7B	FLAN-T5	InstructGPT	ChatGPT
Per Dialogue	1.75/9	2.03/9	9/9	8.98/9
All Questions	166/1800	169/1800	1400/1800	1419/1800

4.3 Question Analysis

How diverse are the automatically generated questions? We analyze 1419 unique questions filtered from 1800 generated questions on 200 random CC (Sharma et al., 2018) samples. To visualize the diversity of the questions, we compute the question embedding based on Sentence-BERT (Reimers & Gurevych, 2019) and reduce the dimension to two for visualization via UMAP (McInnes et al., 2018). The resulting visualization is presented in Fig.6.

Our analysis shows that the questions from ChatCaptioner are diverse and cover various perspectives, including high-level information like *"What is the overall mood of this image?"*, object properties like *"What material are the clothes in the closet made of?"*, environmental information like *"What is the sky in the image like?"*, and so on. This suggests that our automatic questioning mechanism is able to generate diverse questions, explore various aspects of the image, and acquire more informative responses.

Can we use other LLMs as questioners? In addition to ChatGPT, we evaluate the automatic questioning capabilities of several other large language models (LLMs), including InstructGPT (Ouyang et al., 2022), GPT-3 (Brown et al., 2020) without instruction finetuning, FLAN-T5 11B (Chung et al., 2022) (11B parameters), and OPT 6.7B (Zhang et al., 2022) (6.7B parameters). Fig.7 showcases the different question-asking abilities given the initial description from BLIP-2. We observe that InstructGPT shows a similar

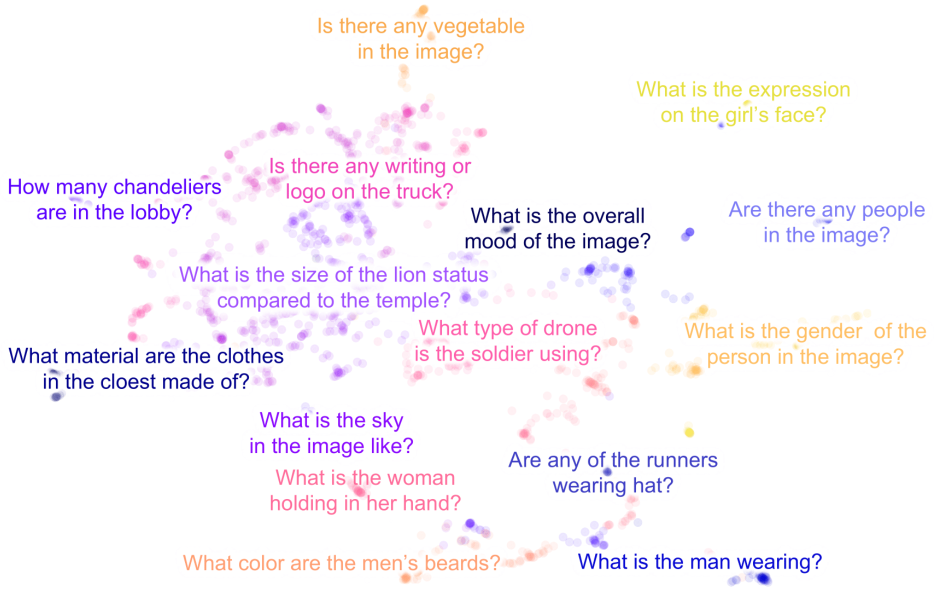
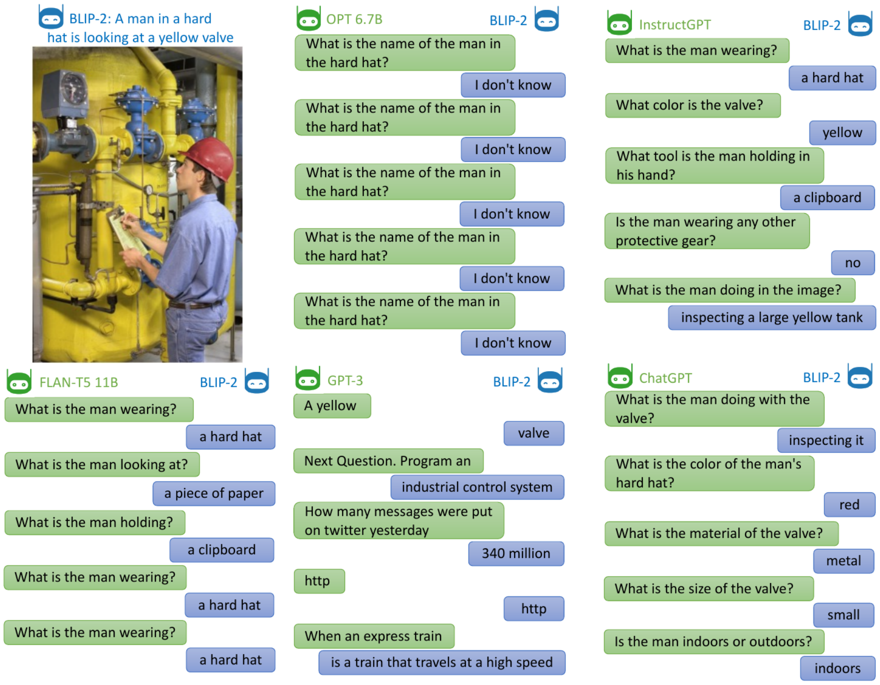


Figure 6: Visualization of question embeddings and question examples printed at the corresponding positions. Our method can ask diverse questions focusing on various perspectives of the image.


BLIP-2: A man in a hard hat is looking at a yellow valve

What is the name of the man in the hard hat?
I don't know
What is the name of the man in the hard hat?
I don't know
What is the name of the man in the hard hat?
I don't know
What is the name of the man in the hard hat?
I don't know
What is the name of the man in the hard hat?
I don't know

What is the man wearing?
a hard hat
What color is the valve?
yellow
What tool is the man holding in his hand?
a clipboard
Is the man wearing any other protective gear?
no
What is the man doing in the image?
inspecting a large yellow tank

What is the man wearing?
a hard hat
What is the man looking at?
a piece of paper
What is the man holding?
a clipboard
What is the man wearing?
a hard hat
What is the man wearing?
a hard hat

A yellow
valve
Next Question. Program an
industrial control system
How many messages were put on twitter yesterday
340 million
http
http
When an express train
is a train that travels at a high speed

What is the man doing with the valve?
inspecting it
What is the color of the man's hard hat?
red
What is the material of the valve?
metal
What is the size of the valve?
small
Is the man indoors or outdoors?
indoors

Figure 7: Question examples from various LLMs. ChatGPT and InstructGPT are able to ask new and informative questions. FLAN-T5 11B and OPT 6.7B tend to repeat old questions. GPT-3 fails to ask any related questions.

question-asking ability as ChatGPT and is able to generate meaningful and diverse questions. While FLAN-T5 11B and OPT 6.7B also exhibits the ability to ask questions, it repeats a single question several times. Interestingly, GPT-3 fails to generate any meaningful questions related to the image.

We further analyze their 1800 generated questions on 200 CC (Sharma et al., 2018) samples. We skip the poorly performed GPT-3 here. LLMs' questioning abilities are evaluated in two ways. Firstly, we measure the average number of unique questions per image, which indicates whether the model can keep asking new questions in a single dialogue. Secondly, we count the total unique questions to see if the questioner could customize the questions according to the given contexts or just ask fixed predefined questions.

Results in Tab.5 reveal that ChatGPT and InstructGPT almost never repeat their question in a single dialogue and generate around 1400 unique questions out of 1800, suggesting that they are able to ask new questions according to the contexts. In contrast, FLAN-T5 11B and OPT 6.7B tend to repeat old questions, averaging about only 2 unique questions per image and generating less than 170 unique questions in total. Our study suggests that to develop a questioning machine that can generate novel and customized questions, it may be necessary to utilize instruction-finetuned LLMs with dozens of billions of parameters.

5 Conclusion

In this work, we discover that advanced large language models possess the ability to pose insightful and diverse questions when provided with well-crafted prompts. Based on our findings, we develop an automatic questioning system named ChatCaptioner for the task of image captioning. By prompting ChatGPT to keep asking questions that expand its understanding of an image, ChatCaptioner guides BLIP-2 to provide comprehensive image information, resulting in image captions that are significantly more detailed and enriched than that from BLIP-2 alone. ChatCaptioner demonstrates the power of automatic questioning systems to effectively extract desired information. Through our work, we aim to draw attention to the potential of automatic questioning systems in AI and inspire further research in various domains.

References

Openai api: Gpt-3.5. `https://platform.openai.com/docs/models/gpt-3-5`, 2023. Accessed: March 7, 2023.

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11569–11579, 2021.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18030–18040, 2022.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

Bilal Ghanem, Lauren Lutz Coleman, Julia Rivard Dexter, Spencer McIntosh von der Ohe, and Alona Fyshe. Question generation for reading comprehension assessment by modeling how and what to ask. *arXiv preprint arXiv:2204.02908*, 2022.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Michael Heilman and Noah A Smith. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 609–617, 2010.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Unnat Jain, Ziyu Zhang, and Alexander G Schwing. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6485–6494, 2017.

Xin Jia, Wenjie Zhou, Xu Sun, and Yunfang Wu. How to ask good questions? try to leverage paraphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6130–6140, 2020.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 workshop on automatic summarization*, pp. 45–51, 2002.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Bang Liu, Mingjun Zhao, Di Niu, Kunfeng Lai, Yancheng He, Haojie Wei, and Yu Xu. Learning to generate questions by learning what not to generate. In *The world wide web conference*, pp. 1106–1118, 2019.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1802–1813, 2016.

Jack Mostow and Wei Chen. Generating instruction automatically for the reading strategy of self-questioning. In *AIED*, pp. 465–472, 2009.

OpenAI. Introducing chatgpt. *https://openai.com/blog/chatgpt*, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Badri Patro, Vinod Kurmi, Sandeep Kumar, and Vinay Namboodiri. Deep bayesian network for visual question generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

Badri Narayana Patro, Sandeep Kumar, Vinod Kumar Kurmi, and Vinay Namboodiri. Multimodal differential network for visual question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4002–4012, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1434. URL `https://aclanthology.org/D18-1434`.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 647–664. Springer, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Tingke Shen, Amlan Kar, and Sanja Fidler. Learning to caption images through a lifetime by asking questions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10393–10402, 2019.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34: 200–212, 2021.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao, and Lucia Specia. Guiding visual question generation. *arXiv preprint arXiv:2110.08226*, 2021.

Qifan Wang, Li Yang, Xiaojun Quan, Fuli Feng, Dongfang Liu, Zenglin Xu, Sinong Wang, and Hao Ma. Learning to generate question by asking question: A primal-dual approach with uncommon word generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 46–61, 2022a.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2020.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022a.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022b.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. *arXiv preprint arXiv:2001.11314*, 2020.

Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2013–2018, 2015.

Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. Automatic generation of grounded visual questions. *arXiv preprint arXiv:1612.06530*, 2016.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Appendix

A.1 Cost

Our method is based on the ChatGPT model, specifically the *gpt-3.5-turbo* version which we access through OpenAI's API. At the time of our project, the cost for using 1000 tokens in *gpt-3.5-turbo* was 0.002 US Dollars. On average, we spent approximately 2500 tokens for each image for ten Question-Answer rounds, which translates to a cost of approximately 0.005 US Dollars per image.

A.2 Yes/No Question Ablation

Usually, yes/no questions contain relatively less information. To reduce the generation of yes/no questions from ChatGPT, we explicitly add a prompt *"Avoid asking yes/no questions"* in the task instruction ρ_{taskQ} and the question instruction ρ_q. Our ablation study in Tab.6 shows that this prompt reduces the generation of yes/no questions from 33% of the cases to 2% in 1800 questions on 200 random CC Sharma et al. (2018) samples, verifying its effectiveness.

Table 6: Effectiveness of the yes/no prompt.

	Total Question	Yes/No Question w/o Prompt	Yes/No Question with Prompt
Num.	1800	595	38
Ratio	-	33%	2%

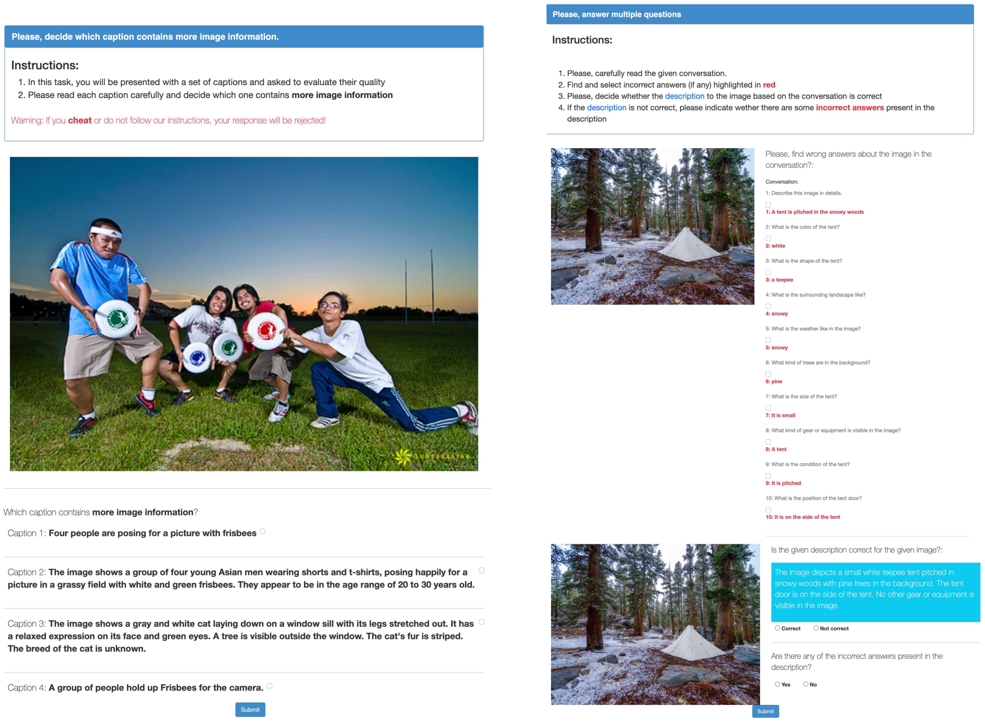
A.3 Using Vicuna as Questioners

In addition to the open-sourced LLMs Flan-T5 11B and OPT 6.7B in the main paper, here we conduct experiments with one of the latest SOTA open-sourced LLMs, Vicuna 13B Chiang et al. (2023), as the questioner and evaluate the generated question quality as shown in Tab.7. Experimental results suggest that Vicuna is able to generate diverse questions like ChatGPT and rarely repeats the questions in a single 9-question dialogue.

Table 7: Number of unique questions.

Unique Q/Total Q	OPT 6.7B	FLAN-T5	Vicuna	ChatGPT
Per Dialogue	1.75/9	2.03/9	8.88/9	8.98/9
All Questions	166/1800	169/1800	1710/1800	1419/1800

A.4 Human Evaluation Interface



Please, decide which caption contains more image information.

Instructions:

1. In this task, you will be presented with a set of captions and asked to evaluate their quality
2. Please read each caption carefully and decide which one contains **more image information**

Warning: If you **cheat** or do not follow our instructions, your response will be rejected!

Which caption contains **more image information**?

Caption 1: **Four people are posing for a picture with frisbees** ○

Caption 2: **The image shows a group of four young Asian men wearing shorts and t-shirts, posing happily for a picture in a grassy field with white and green frisbees. They appear to be in the age range of 20 to 30 years old.** ○

Caption 3: **The image shows a gray and white cat laying down on a window sill with its legs stretched out. It has a relaxed expression on its face and green eyes. A tree is visible outside the window. The cat's fur is striped. The breed of the cat is unknown.** ○

Caption 4: **A group of people hold up Frisbees for the camera.** ○

Please, answer multiple questions

Instructions:

1. Please, carefully read the given conversation.
2. Find and select incorrect answers (if any) highlighted in **red**
3. Please, decide whether the description to the image based on the conversation is correct
4. If the description is not correct, please indicate wether there are some **incorrect answers** present in the description

Please, find wrong answers about the image in the conversation?:

Conversation:

1: Describe this image in details.

☐ **1: A tent is pitched in the snowy woods**

2: What is the color of the tent?

☐ **2: white**

3: What is the shape of the tent?

☐ **3: a teepee**

4: What is the surrounding landscape like?

☐ **4: snowy**

5: What is the weather like in the image?

☐ **5: snowy**

6: What kind of trees are in the background?

☐ **6: pine**

7: What is the size of the tent?

☐ **7: It is small**

8: What kind of gear or equipment is visible in the image?

☐ **8: A tent**

9: What is the condition of the tent?

☐ **9: It is pitched**

10: What is the position of the tent door?

☐ **10: It is on the side of the tent**

Is the given description correct for the given image?:

The image depicts a small white teepee tent pitched in snowy woods with pine trees in the background. The tent door is on the side of the tent. No other gear or equipment is visible in the image.

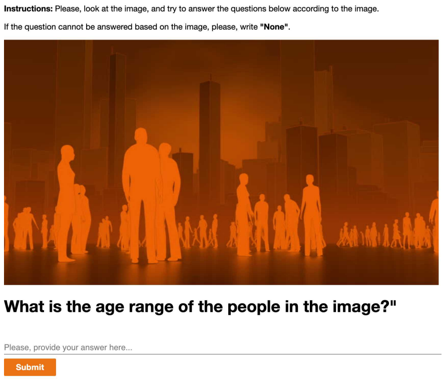
○ **Correct** ○ **Not correct**

Are there any of the incorrect answers present in the description?

○ **Yes** ○ **No**

(a) Human evaluation interface of the information experiments.

(b) Human evaluation interface of the correctness experiments.



Instructions: Please, look at the image, and try to answer the questions below according to the image.

If the question cannot be answered based on the image, please, write **"None"**.

What is the age range of the people in the image?"

Please, provide your answer here...

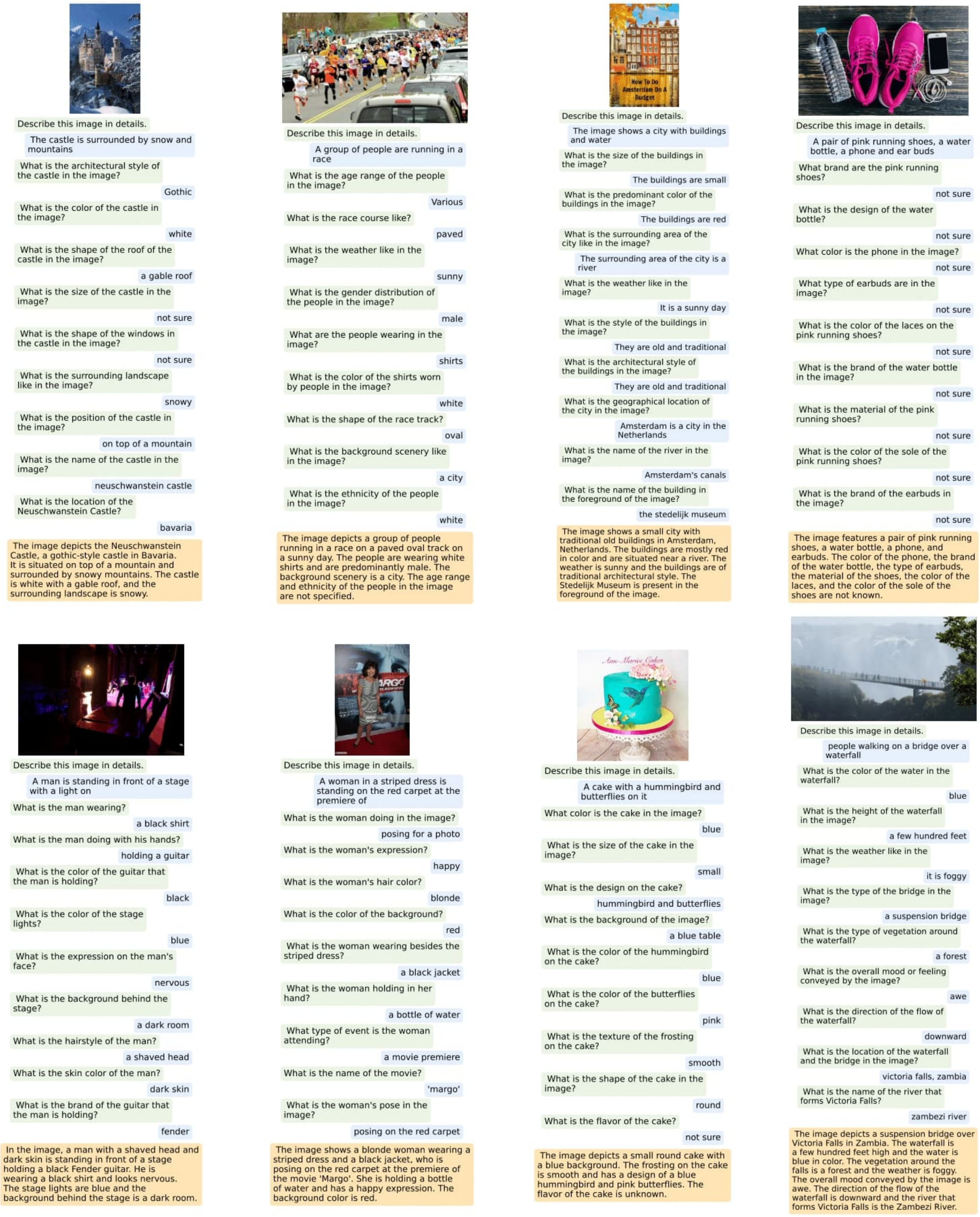
(c) Human evaluation interface of the question-answerable experiments.

Figure 8: Human evaluation interfaces

A.5 Random Examples of Generated Questions in Conceptual Captioning Dataset

What is the material of the pier in the image?
What is the position of the sign that says "No Swimming Allowed" on the dock?
What is the material of the valve?
What is the color of the plate on which the cake is placed?
What is the expression on the man’s face?
What is the boy’s expression while playing with the sailboat?
What is the angle of the camera in the image?
What type of flower is it?
What is the color of the woman’s glasses?
What is written on the quote on the mountain?
What is the design on the cake?
What is the woman’s hair color in the image?
Are the man and woman standing or sitting in the image?
What is the location of the scene depicted in the image?
What is the boy’s expression?
What is the material of the pink running shoes?
What is the expression on the man’s face?
What type of vegetation surrounds the pond in the image?
What is the size of the fountain in the image?
What is the name of the mountain range seen in the background of the image?
What is the name of the park?
What is the design of the woman’s dress?
What is the color of the chainsaw?
What is the ethnicity of the two men in the image?
What is the woman’s pose in the photo?
What modifications, if any, have been made to the car in the image?
What kind of donuts are in the box?
What is the woman’s age range in the image?
What is the weather like in the image?
What is the man’s posture like in the image?
What kind of lighting is in the room?
What is the woman’s hair color in the image?
What is the woman wearing in the image?
What is the woman’s pose in the image?
What is the type of the lightning bolt? (e.g. cloud-to-ground, cloud-to-cloud)
What is the context or setting of the image?
What type of event is taking place where the man is performing?
What is the shape of the base or stand of the ice sculpture?
What is the size of the tent?
What is the boy wearing?
What is the man doing?
What is the material of the sink in the image?
What is the age range of the people in the image?
What is the style of the hat?
What is the background of the image?
What is the surrounding environment like in the image?
What is the woman wearing?
What is the shape of the backyard?
What is the material of the backpack the man is carrying?
What is the texture of the granny square?
What is the style of the bridge?
What is the size of the sun in the image?

A.6 More Qualitative Results



Describe this image in details.
The castle is surrounded by snow and mountains
What is the architectural style of the castle in the image?
Gothic
What is the color of the castle in the image?
white
What is the shape of the roof of the castle in the image?
a gable roof
What is the size of the castle in the image?
not sure
What is the shape of the windows in the castle in the image?
not sure
What is the surrounding landscape like in the image?
snowy
What is the position of the castle in the image?
on top of a mountain
What is the name of the castle in the image?
neuschwanstein castle
What is the location of the Neuschwanstein Castle?
bavaria
The image depicts the Neuschwanstein Castle, a gothic-style castle in Bavaria. It is situated on top of a mountain and surrounded by snowy mountains. The castle is white with a gable roof, and the surrounding landscape is snowy.

Describe this image in details.
A group of people are running in a race
What is the age range of the people in the image?
Various
What is the race course like?
paved
What is the weather like in the image?
sunny
What is the gender distribution of the people in the image?
male
What are the people wearing in the image?
shirts
What is the color of the shirts worn by people in the image?
white
What is the shape of the race track?
oval
What is the background scenery like in the image?
a city
What is the ethnicity of the people in the image?
white
The image depicts a group of people running in a race on a paved oval track on a sunny day. The people are wearing white shirts and are predominantly male. The background scenery is a city. The age range and ethnicity of the people in the image are not specified.

Describe this image in details.
The image shows a city with buildings and water
What is the size of the buildings in the image?
The buildings are small
What is the predominant color of the buildings in the image?
The buildings are red
What is the surrounding area of the city like in the image?
The surrounding area of the city is a river
What is the weather like in the image?
It is a sunny day
What is the style of the buildings in the image?
They are old and traditional
What is the architectural style of the buildings in the image?
They are old and traditional
What is the geographical location of the city in the image?
Amsterdam is a city in the Netherlands
What is the name of the river in the image?
Amsterdam's canals
What is the name of the building in the foreground of the image?
the stedelijk museum
The image shows a small city with traditional old buildings in Amsterdam, Netherlands. The buildings are mostly red in color and are situated near a river. The weather is sunny and the buildings are of traditional architectural style. The Stedelijk Museum is present in the foreground of the image.

Describe this image in details.
A pair of pink running shoes, a water bottle, a phone and ear buds
What brand are the pink running shoes?
not sure
What is the design of the water bottle?
not sure
What color is the phone in the image?
not sure
What type of earbuds are in the image?
not sure
What is the color of the laces on the pink running shoes?
not sure
What is the brand of the water bottle in the image?
not sure
What is the material of the pink running shoes?
not sure
What is the color of the sole of the pink running shoes?
not sure
What is the brand of the earbuds in the image?
not sure
The image features a pair of pink running shoes, a water bottle, a phone, and earbuds. The color of the phone, the brand of the water bottle, the type of earbuds, the material of the shoes, the color of the laces, and the color of the sole of the shoes are not known.

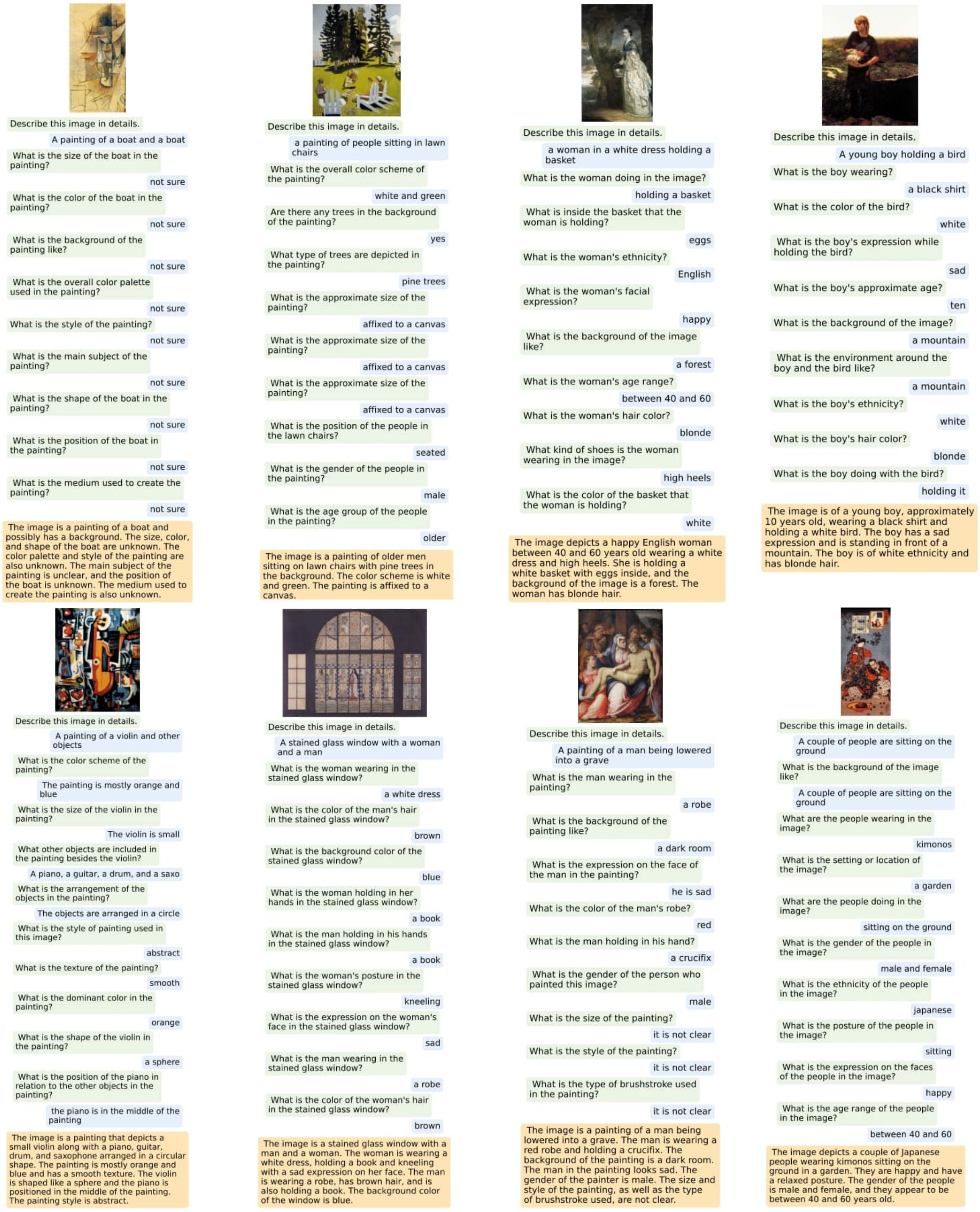
Describe this image in details.
A man is standing in front of a stage with a light on
What is the man wearing?
a black shirt
What is the man doing with his hands?
holding a guitar
What is the color of the guitar that the man is holding?
black
What is the color of the stage lights?
blue
What is the expression on the man's face?
nervous
What is the background behind the stage?
a dark room
What is the hairstyle of the man?
a shaved head
What is the skin color of the man?
dark skin
What is the brand of the guitar that the man is holding?
fender
In the image, a man with a shaved head and dark skin is standing in front of a stage holding a black Fender guitar. He is wearing a black shirt and looks nervous. The stage lights are blue and the background behind the stage is a dark room.

Describe this image in details.
A woman in a striped dress is standing on the red carpet at the premiere of
What is the woman doing in the image?
posing for a photo
What is the woman's expression?
happy
What is the woman's hair color?
blonde
What is the color of the background?
red
What is the woman wearing besides the striped dress?
a black jacket
What is the woman holding in her hand?
a bottle of water
What type of event is the woman attending?
a movie premiere
What is the name of the movie?
'margo'
What is the woman's pose in the image?
posing on the red carpet
The image shows a blonde woman wearing a striped dress and a black jacket, who is posing on the red carpet at the premiere of the movie 'Margo'. She is holding a bottle of water and has a happy expression. The background color is red.

Describe this image in details.
A cake with a hummingbird and butterflies on it
What color is the cake in the image?
blue
What is the size of the cake in the image?
small
What is the design on the cake?
hummingbird and butterflies
What is the background of the image?
a blue table
What is the color of the hummingbird on the cake?
blue
What is the color of the butterflies on the cake?
pink
What is the texture of the frosting on the cake?
smooth
What is the shape of the cake in the image?
round
What is the flavor of the cake?
not sure
The image depicts a small round cake with a blue background. The frosting on the cake is smooth and has a design of a blue hummingbird and pink butterflies. The flavor of the cake is unknown.

Describe this image in details.
people walking on a bridge over a waterfall
What is the color of the water in the waterfall?
blue
What is the height of the waterfall in the image?
a few hundred feet
What is the weather like in the image?
it is foggy
What is the type of the bridge in the image?
a suspension bridge
What is the type of vegetation around the waterfall?
a forest
What is the overall mood or feeling conveyed by the image?
awe
What is the direction of the flow of the waterfall?
downward
What is the location of the waterfall and the bridge in the image?
victoria falls, zambia
What is the name of the river that forms Victoria Falls?
zambezi river
The image depicts a suspension bridge over Victoria Falls in Zambia. The waterfall is a few hundred feet high and the water is blue in color. The vegetation around the falls is a forest and the weather is foggy. The overall mood conveyed by the image is awe. The direction of the flow of the waterfall is downward and the river that forms Victoria Falls is the Zambezi River.

Figure 9: Random samples from CC Sharma et al. (2018).



Describe this image in details.

A painting of a boat and a boat

What is the size of the boat in the painting?

not sure

What is the color of the boat in the painting?

not sure

What is the background of the painting like?

not sure

What is the overall color palette used in the painting?

not sure

What is the style of the painting?

not sure

What is the main subject of the painting?

not sure

What is the shape of the boat in the painting?

not sure

What is the position of the boat in the painting?

not sure

What is the medium used to create the painting?

not sure

The image is a painting of a boat and possibly has a background. The size, color, and shape of the boat are unknown. The color palette and style of the painting are also unknown. The main subject of the painting is unclear, and the position of the boat is unknown. The medium used to create the painting is also unknown.

Describe this image in details.

a painting of people sitting in lawn chairs

What is the overall color scheme of the painting?

white and green

Are there any trees in the background of the painting?

yes

What type of trees are depicted in the painting?

pine trees

What is the approximate size of the painting?

affixed to a canvas

What is the approximate size of the painting?

affixed to a canvas

What is the approximate size of the painting?

affixed to a canvas

What is the position of the people in the lawn chairs?

seated

What is the gender of the people in the painting?

male

What is the age group of the people in the painting?

older

The image is a painting of older men sitting on lawn chairs with pine trees in the background. The color scheme is white and green. The painting is affixed to a canvas.

Describe this image in details.

a woman in a white dress holding a basket

What is the woman doing in the image?

holding a basket

What is inside the basket that the woman is holding?

eggs

What is the woman's ethnicity?

English

What is the woman's facial expression?

happy

What is the background of the image like?

a forest

What is the woman's age range?

between 40 and 60

What is the woman's hair color?

blonde

What kind of shoes is the woman wearing in the image?

high heels

What is the color of the basket that the woman is holding?

white

The image depicts a happy English woman between 40 and 60 years old wearing a white dress and high heels. She is holding a white basket with eggs inside, and the background of the image is a forest. The woman has blonde hair.

Describe this image in details.

A young boy holding a bird

What is the boy wearing?

a black shirt

What is the color of the bird?

white

What is the boy's expression while holding the bird?

sad

What is the boy's approximate age?

ten

What is the background of the image?

a mountain

What is the environment around the boy and the bird like?

a mountain

What is the boy's ethnicity?

white

What is the boy's hair color?

blonde

What is the boy doing with the bird?

holding it

The image is of a young boy, approximately 10 years old, wearing a black shirt and holding a white bird. The boy has a sad expression and is standing in front of a mountain. The boy is of white ethnicity and has blonde hair.

Describe this image in details.

A painting of a violin and other objects

What is the color scheme of the painting?

The painting is mostly orange and blue

What is the size of the violin in the painting?

The violin is small

What other objects are included in the painting besides the violin?

A piano, a guitar, a drum, and a saxo

What is the arrangement of the objects in the painting?

The objects are arranged in a circle

What is the style of painting used in this image?

abstract

What is the texture of the painting?

smooth

What is the dominant color in the painting?

orange

What is the shape of the violin in the painting?

a sphere

What is the position of the piano in relation to the other objects in the painting?

the piano is in the middle of the painting

The image is a painting that depicts a small violin along with a piano, guitar, drum, and saxophone arranged in a circular shape. The painting is mostly orange and blue and has a smooth texture. The violin is shaped like a sphere and the piano is positioned in the middle of the painting. The painting style is abstract.

Describe this image in details.

A stained glass window with a woman and a man

What is the woman wearing in the stained glass window?

a white dress

What is the color of the man's hair in the stained glass window?

brown

What is the background color of the stained glass window?

blue

What is the woman holding in her hands in the stained glass window?

a book

What is the man holding in his hands in the stained glass window?

a book

What is the woman's posture in the stained glass window?

kneeling

What is the expression on the woman's face in the stained glass window?

sad

What is the man wearing in the stained glass window?

a robe

What is the color of the woman's hair in the stained glass window?

brown

The image is a stained glass window with a man and a woman. The woman is wearing a white dress, holding a book and kneeling with a sad expression on her face. The man is wearing a robe, has brown hair, and is also holding a book. The background color of the window is blue.

Describe this image in details.

A painting of a man being lowered into a grave

What is the man wearing in the painting?

a robe

What is the background of the painting like?

a dark room

What is the expression on the face of the man in the painting?

he is sad

What is the color of the man's robe?

red

What is the man holding in his hand?

a crucifix

What is the gender of the person who painted this image?

male

What is the size of the painting?

it is not clear

What is the style of the painting?

it is not clear

What is the type of brushstroke used in the painting?

it is not clear

The image is a painting of a man being lowered into a grave. The man is wearing a red robe and holding a crucifix. The background of the painting is a dark room. The man in the painting looks sad. The gender of the painter is male. The size and style of the painting, as well as the type of brushstroke used, are not clear.

Describe this image in details.

A couple of people are sitting on the ground

What is the background of the image like?

A couple of people are sitting on the ground

What are the people wearing in the image?

kimonos

What is the setting or location of the image?

a garden

What are the people doing in the image?

sitting on the ground

What is the gender of the people in the image?

male and female

What is the ethnicity of the people in the image?

japanese

What is the posture of the people in the image?

sitting

What is the expression on the faces of the people in the image?

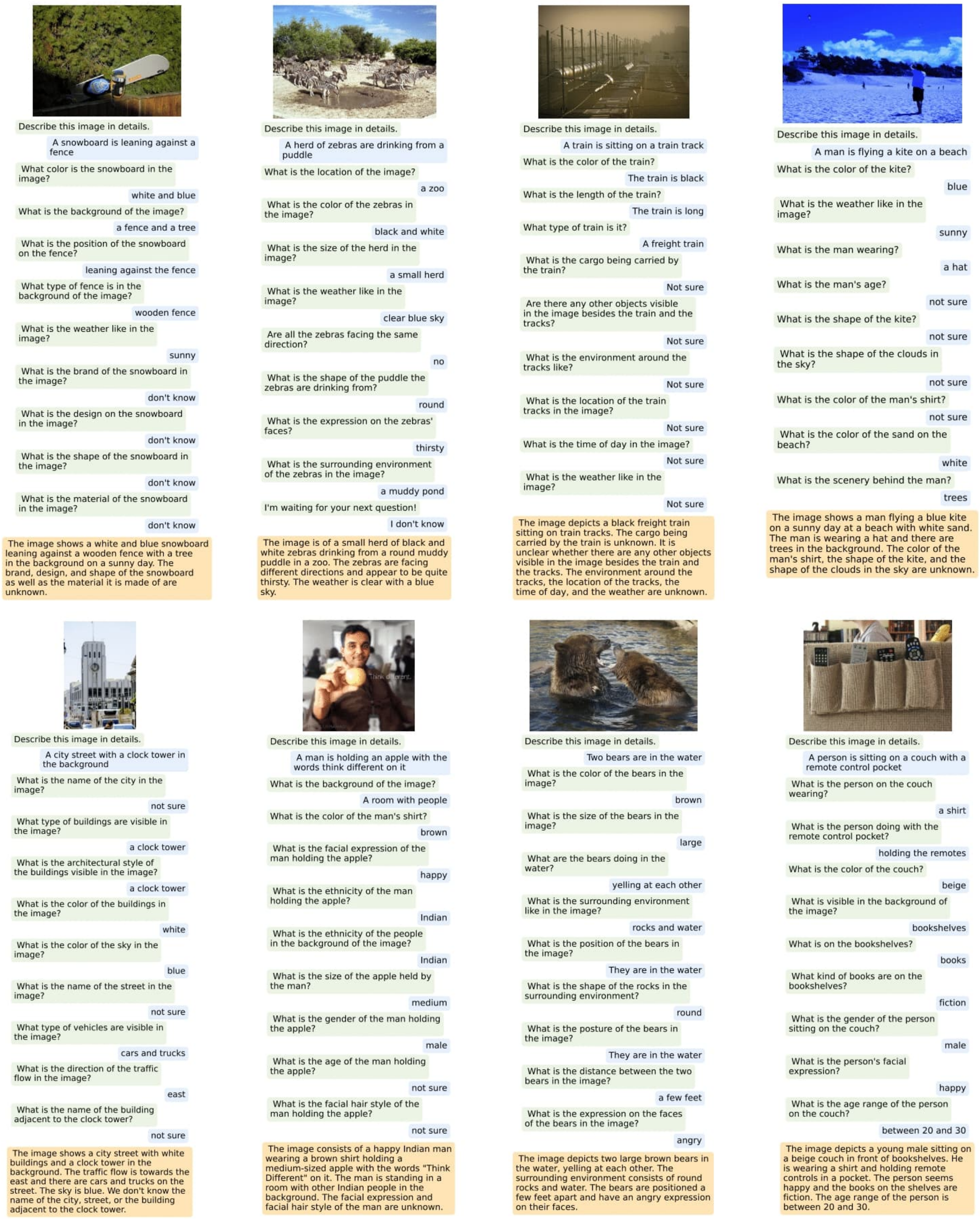
happy

What is the age range of the people in the image?

between 40 and 60

The image depicts a couple of Japanese people wearing kimonos sitting on the ground in a garden. They are happy and have a relaxed posture. The gender of the people is male and female, and they appear to be between 40 and 60 years old.

Figure 10: Random samples from WikiArt Saleh & Elgammal (2015).



Describe this image in details.
A snowboard is leaning against a fence
What color is the snowboard in the image?
white and blue
What is the background of the image?
a fence and a tree
What is the position of the snowboard on the fence?
leaning against the fence
What type of fence is in the background of the image?
wooden fence
What is the weather like in the image?
sunny
What is the brand of the snowboard in the image?
don't know
What is the design on the snowboard in the image?
don't know
What is the shape of the snowboard in the image?
don't know
What is the material of the snowboard in the image?
don't know
The image shows a white and blue snowboard leaning against a wooden fence with a tree in the background on a sunny day. The brand, design, and shape of the snowboard as well as the material it is made of are unknown.

Describe this image in details.
A herd of zebras are drinking from a puddle
What is the location of the image?
a zoo
What is the color of the zebras in the image?
black and white
What is the size of the herd in the image?
a small herd
What is the weather like in the image?
clear blue sky
Are all the zebras facing the same direction?
no
What is the shape of the puddle the zebras are drinking from?
round
What is the expression on the zebras' faces?
thirsty
What is the surrounding environment of the zebras in the image?
a muddy pond
I'm waiting for your next question!
I don't know
The image is of a small herd of black and white zebras drinking from a round muddy puddle in a zoo. The zebras are facing different directions and appear to be quite thirsty. The weather is clear with a blue sky.

Describe this image in details.
A train is sitting on a train track
What is the color of the train?
The train is black
What is the length of the train?
The train is long
What type of train is it?
A freight train
What is the cargo being carried by the train?
Not sure
Are there any other objects visible in the image besides the train and the tracks?
Not sure
What is the environment around the tracks like?
Not sure
What is the location of the train tracks in the image?
Not sure
What is the time of day in the image?
Not sure
What is the weather like in the image?
Not sure
The image depicts a black freight train sitting on train tracks. The cargo being carried by the train is unknown. It is unclear whether there are any other objects visible in the image besides the train and the tracks. The environment around the tracks, the location of the tracks, the time of day, and the weather are unknown.

Describe this image in details.
A man is flying a kite on a beach
What is the color of the kite?
blue
What is the weather like in the image?
sunny
What is the man wearing?
a hat
What is the man's age?
not sure
What is the shape of the kite?
not sure
What is the shape of the clouds in the sky?
not sure
What is the color of the man's shirt?
not sure
What is the color of the sand on the beach?
white
What is the scenery behind the man?
trees
The image shows a man flying a blue kite on a sunny day at a beach with white sand. The man is wearing a hat and there are trees in the background. The color of the man's shirt, the shape of the kite, and the shape of the clouds in the sky are unknown.

Describe this image in details.
A city street with a clock tower in the background
What is the name of the city in the image?
not sure
What type of buildings are visible in the image?
a clock tower
What is the architectural style of the buildings visible in the image?
a clock tower
What is the color of the buildings in the image?
white
What is the color of the sky in the image?
blue
What is the name of the street in the image?
not sure
What type of vehicles are visible in the image?
cars and trucks
What is the direction of the traffic flow in the image?
east
What is the name of the building adjacent to the clock tower?
not sure
The image shows a city street with white buildings and a clock tower in the background. The traffic flow is towards the east and there are cars and trucks on the street. The sky is blue. We don't know the name of the city, street, or the building adjacent to the clock tower.

Describe this image in details.
A man is holding an apple with the words think different on it
What is the background of the image?
A room with people
What is the color of the man's shirt?
brown
What is the facial expression of the man holding the apple?
happy
What is the ethnicity of the man holding the apple?
Indian
What is the ethnicity of the people in the background of the image?
Indian
What is the size of the apple held by the man?
medium
What is the gender of the man holding the apple?
male
What is the age of the man holding the apple?
not sure
What is the facial hair style of the man holding the apple?
not sure
The image consists of a happy Indian man wearing a brown shirt holding a medium-sized apple with the words "Think Different" on it. The man is standing in a room with other Indian people in the background. The facial expression and facial hair style of the man are unknown.

Describe this image in details.
Two bears are in the water
What is the color of the bears in the image?
brown
What is the size of the bears in the image?
large
What are the bears doing in the water?
yelling at each other
What is the surrounding environment like in the image?
rocks and water
What is the position of the bears in the image?
They are in the water
What is the shape of the rocks in the surrounding environment?
round
What is the posture of the bears in the image?
They are in the water
What is the distance between the two bears in the image?
a few feet
What is the expression on the faces of the bears in the image?
angry
The image depicts two large brown bears in the water, yelling at each other. The surrounding environment consists of round rocks and water. The bears are positioned a few feet apart and have an angry expression on their faces.

Describe this image in details.
A person is sitting on a couch with a remote control pocket
What is the person on the couch wearing?
a shirt
What is the person doing with the remote control pocket?
holding the remotes
What is the color of the couch?
beige
What is visible in the background of the image?
bookshelves
What is on the bookshelves?
books
What kind of books are on the bookshelves?
fiction
What is the gender of the person sitting on the couch?
male
What is the person's facial expression?
happy
What is the age range of the person on the couch?
between 20 and 30
The image depicts a young male sitting on a beige couch in front of bookshelves. He is wearing a shirt and holding remote controls in a pocket. The person seems happy and the books on the shelves are fiction. The age range of the person is between 20 and 30.

Figure 11: Random samples from COCO Lin et al. (2014).