InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning (NeurIPS 2023)

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, Steven Hoi

Presented by: Selina Cheng

Outline

1. Background and Overview
2. Methods/Architecture
3. Experimental Design
4. Training/Evaluation
5. Results
6. Limitations
7. Takeaway

Background

- One aspiration of AI research: build a model that can solve **any** arbitrary task
- Building general-purpose vision-language models (VLM) is challenging
  - Rich input distributions
  - Task diversity
- We want broad generalization to vision-language tasks

★ “*A simple caption or description of the image may not be sufficient to accurately answer queries because not all of the necessary details may be included.*”

Before InstructBLIP

- General-purpose language models: Large-scale pretraining and instruction tuning effective
- Vision-language pretraining has been widely studied. Vision-language instruction tuning?
  - Not so much before InstructBLIP.
  - This is a study on *both methods combined*.

Overview of InstructBLIP

- InstructBLIP (open-source):
    - **Built on** the work of **BLIP** and **BLIP-2**
    - Pretrained BLIP-2 models
        - Adapt frozen instruction-tuned LLMs for visual input
    - Instruction-aware Query Transformer (Q-Former)
        - Extract most useful visual features specific to prompt
            - → **dynamic** visual representations fed into the LLM
        - Advantageous if task instructions significantly vary for same input image

Methods

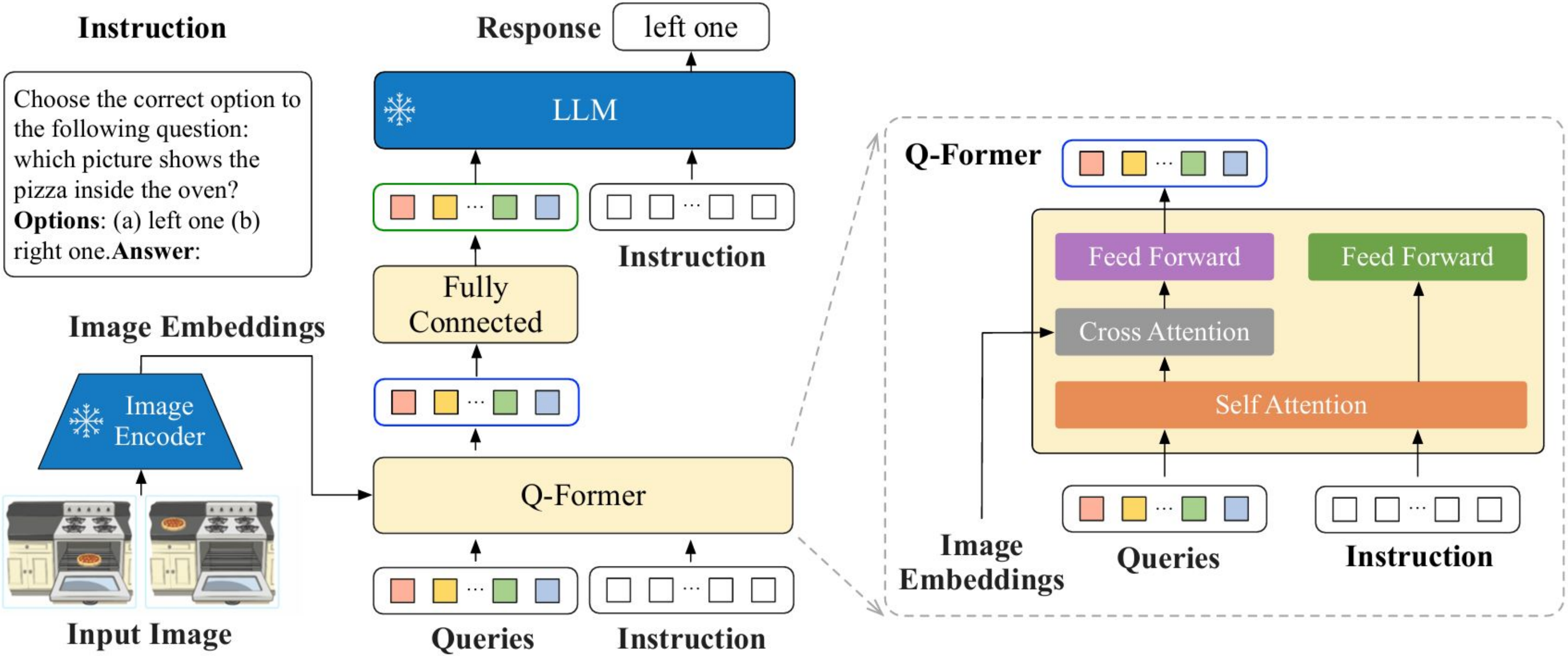
- Initialize training with pre-trained BLIP-2 model:
  - Image encoder
  - LLM
  - Q-Former (**bridge**)

Pretraining Q-Former in two stages

1. Prime the LLM to handle vision-language (VL) input with frozen image encoder
2. Adapt Q-Former instruction-aware visual embeddings into soft prompts
   - Guiding the LLM to give an answer reflecting both image and instruction

Finetuning Q-Former with instruction tuning

- Keep image encoder and LLM **frozen**
- Self-attention layer: refine query embeddings through shared context
- Cross-attention layer: extract task-relevant information from image
- Feed-forward network layer: refine query token representations
- Projection layer: transform final query embeddings into soft visual prompts
- Feed soft prompts into LLM to generate output



Instruction

Choose the correct option to the following question: which picture shows the pizza inside the oven? **Options**: (a) left one (b) right one.**Answer**:

Response

left one

LLM

Instruction

Fully Connected

Image Embeddings

Image Encoder

Q-Former

Input Image

Queries

Instruction

Q-Former

Feed Forward

Feed Forward

Cross Attention

Self Attention

Image Embeddings

Queries

Instruction

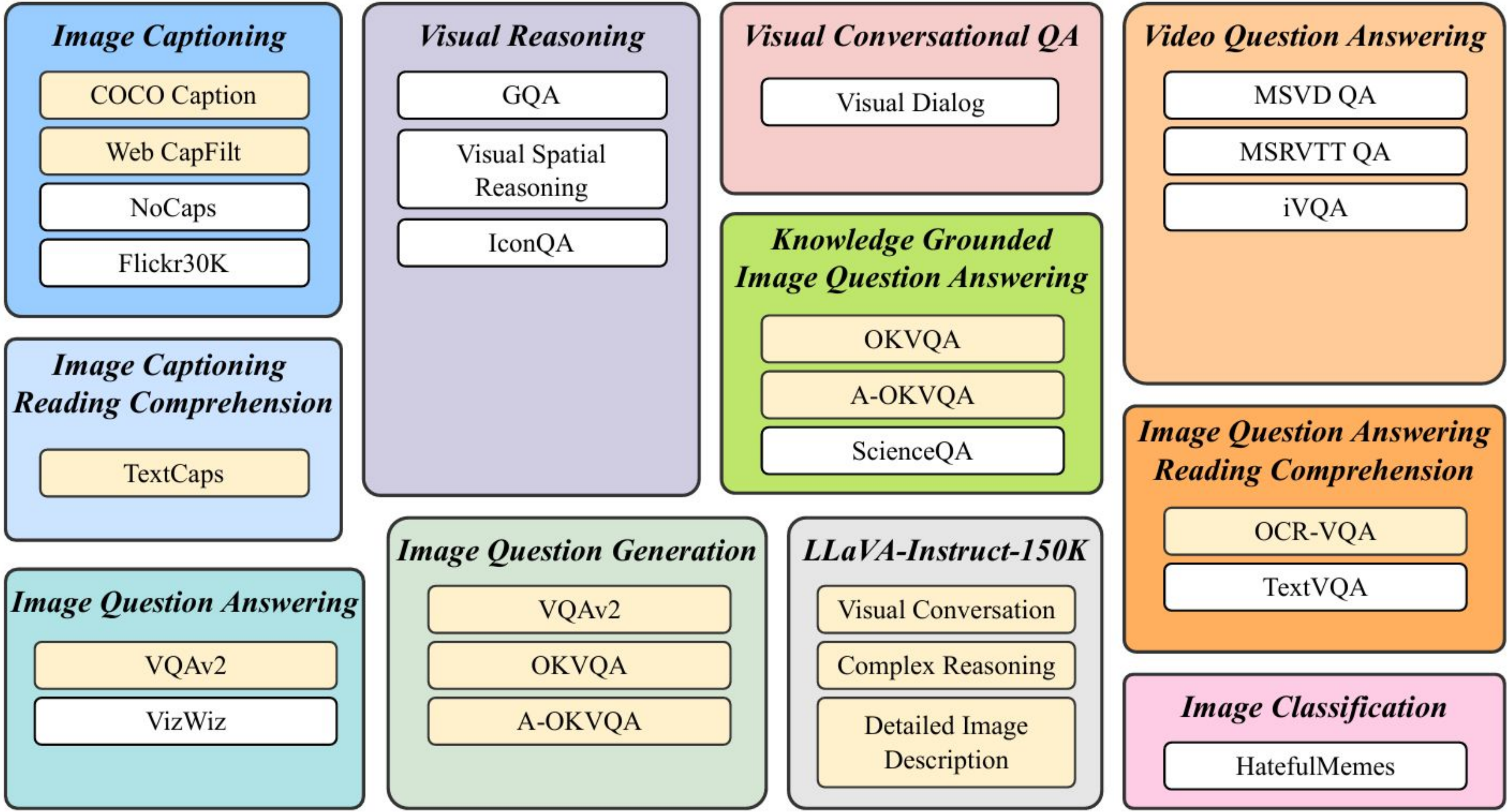


Image Captioning

- COCO Caption
- Web CapFilt
- NoCaps
- Flickr30K

Visual Reasoning

- GQA
- Visual Spatial Reasoning
- IconQA

Visual Conversational QA

- Visual Dialog

Video Question Answering

- MSVD QA
- MSRVTT QA
- iVQA

Knowledge Grounded Image Question Answering

- OKVQA
- A-OKVQA
- ScienceQA

Image Captioning Reading Comprehension

- TextCaps

Image Question Answering Reading Comprehension

- OCR-VQA
- TextVQA

Image Question Answering

- VQAv2
- VizWiz

Image Question Generation

- VQAv2
- OKVQA
- A-OKVQA

LLaVA-Instruct-150K

- Visual Conversation
- Complex Reasoning
- Detailed Image Description

Image Classification

- HatefulMemes

Experimental Design (cont.)

- 4 variations of BLIP-2 with different frozen LLMs
    1. FlanT5-XL (3B)
    2. FlanT5-XXL (11B)
    3. Vicuna-7B
    4. Vicuna-13B

Training and Evaluation

- Training
  - Instruction tuning using training sets (held-in datasets)
  - Validation sets (held-in datasets)
  - Backpropagation/minimization of model loss
- Evaluation (held-out datasets)
  - Evaluate zero-shot performance on unseen data
  - Comparing to previous SOTA models (BLIP-2, Flamingo)

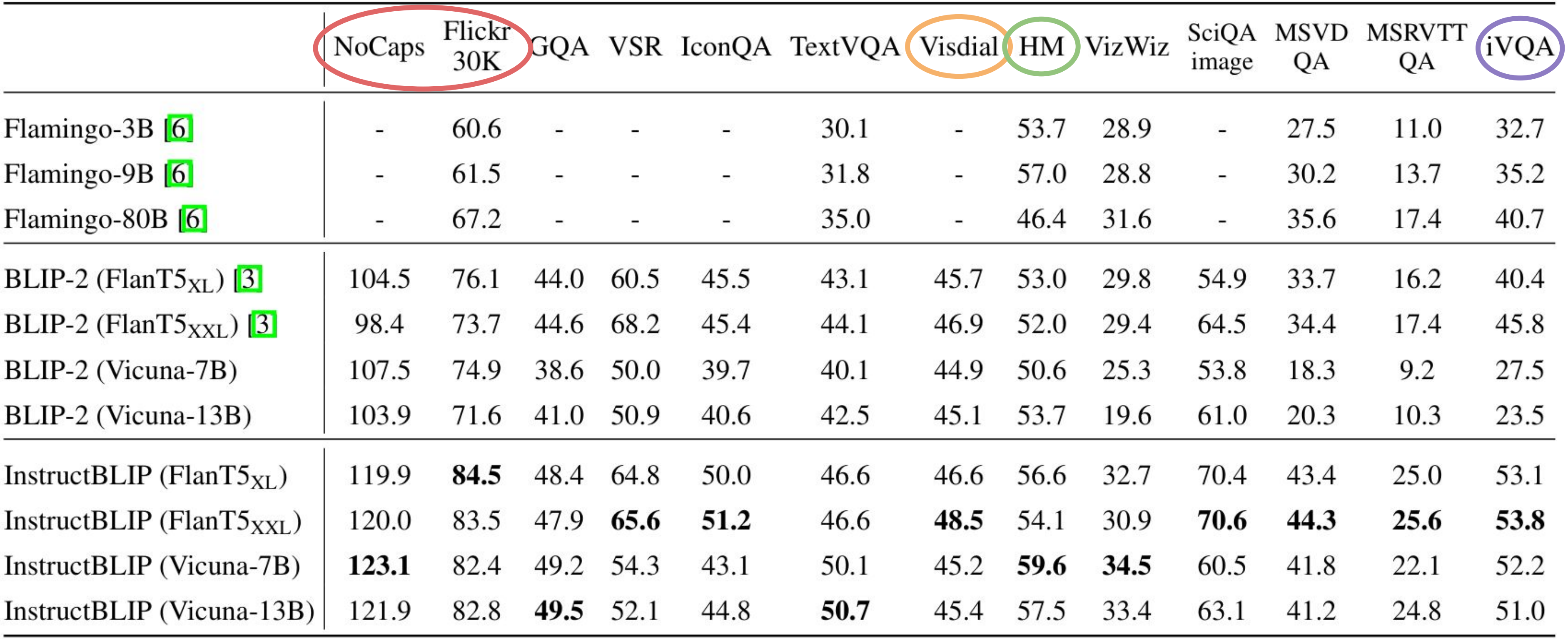
Metrics

- **Top-1 accuracy** (%)
- **CIDEr score** (Consensus-based Image Description Evaluation)
  - ≥100 → high content match
  - Datasets: NoCaps, Flickr 30K
- Mean Reciprocal Rank (**MRR**) for Visual Dialog (Visdial)
  - Rank position of 1st correct answer (0.0-100.0)
- **AUC** (Area Under Curve) **score** for HatefulMemes (HM)
  - Distinguish between pos/neg (0.0, 50.0 = random, 100.0)
- **iVQA** (Inverse Visual QA) **accuracy** (higher is good)

Results

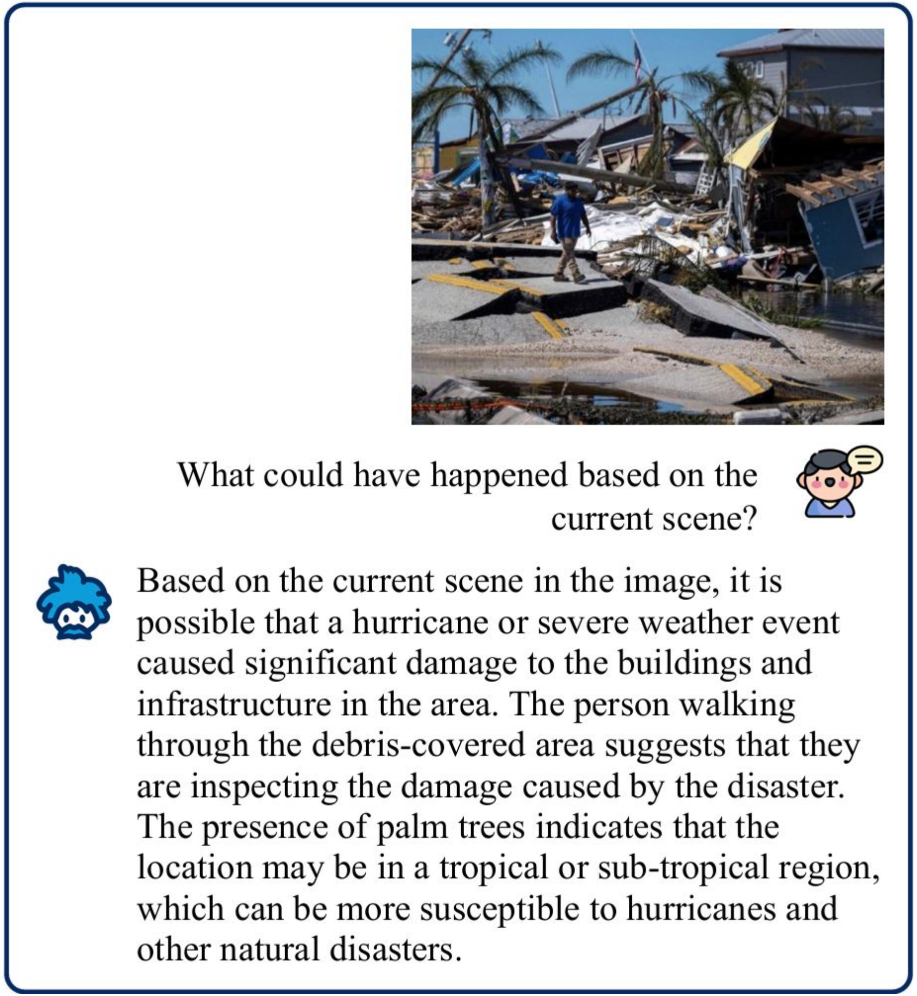
Quantitative Results

- New zero-shot SOTA results on **all datasets**
- Consistently significantly surpasses BLIP-2 across all LLMs
  - Demonstrates effectiveness of instruction tuning
- Up to 47.1% relative improvement on one Video QA dataset over previous SOTA
  - Never trained with temporal video data



	NoCaps	Flickr 30K	GQA	VSR	IconQA	TextVQA	Visdial	HM	VizWiz	SciQA image	MSVD QA	MSRVTT QA	iVQA
Flamingo-3B [6]	-	60.6	-	-	-	30.1	-	53.7	28.9	-	27.5	11.0	32.7
Flamingo-9B [6]	-	61.5	-	-	-	31.8	-	57.0	28.8	-	30.2	13.7	35.2
Flamingo-80B [6]	-	67.2	-	-	-	35.0	-	46.4	31.6	-	35.6	17.4	40.7
BLIP-2 (FlanT5$_{XL}$) [3]	104.5	76.1	44.0	60.5	45.5	43.1	45.7	53.0	29.8	54.9	33.7	16.2	40.4
BLIP-2 (FlanT5$_{XXL}$) [3]	98.4	73.7	44.6	68.2	45.4	44.1	46.9	52.0	29.4	64.5	34.4	17.4	45.8
BLIP-2 (Vicuna-7B)	107.5	74.9	38.6	50.0	39.7	40.1	44.9	50.6	25.3	53.8	18.3	9.2	27.5
BLIP-2 (Vicuna-13B)	103.9	71.6	41.0	50.9	40.6	42.5	45.1	53.7	19.6	61.0	20.3	10.3	23.5
InstructBLIP (FlanT5$_{XL}$)	119.9	**84.5**	48.4	64.8	50.0	46.6	46.6	56.6	32.7	70.4	43.4	25.0	53.1
InstructBLIP (FlanT5$_{XXL}$)	120.0	83.5	47.9	**65.6**	**51.2**	46.6	**48.5**	54.1	30.9	**70.6**	**44.3**	**25.6**	**53.8**
InstructBLIP (Vicuna-7B)	**123.1**	82.4	49.2	54.3	43.1	50.1	45.2	**59.6**	**34.5**	60.5	41.8	22.1	52.2
InstructBLIP (Vicuna-13B)	121.9	82.8	**49.5**	52.1	44.8	**50.7**	45.4	57.5	33.4	63.1	41.2	24.8	51.0

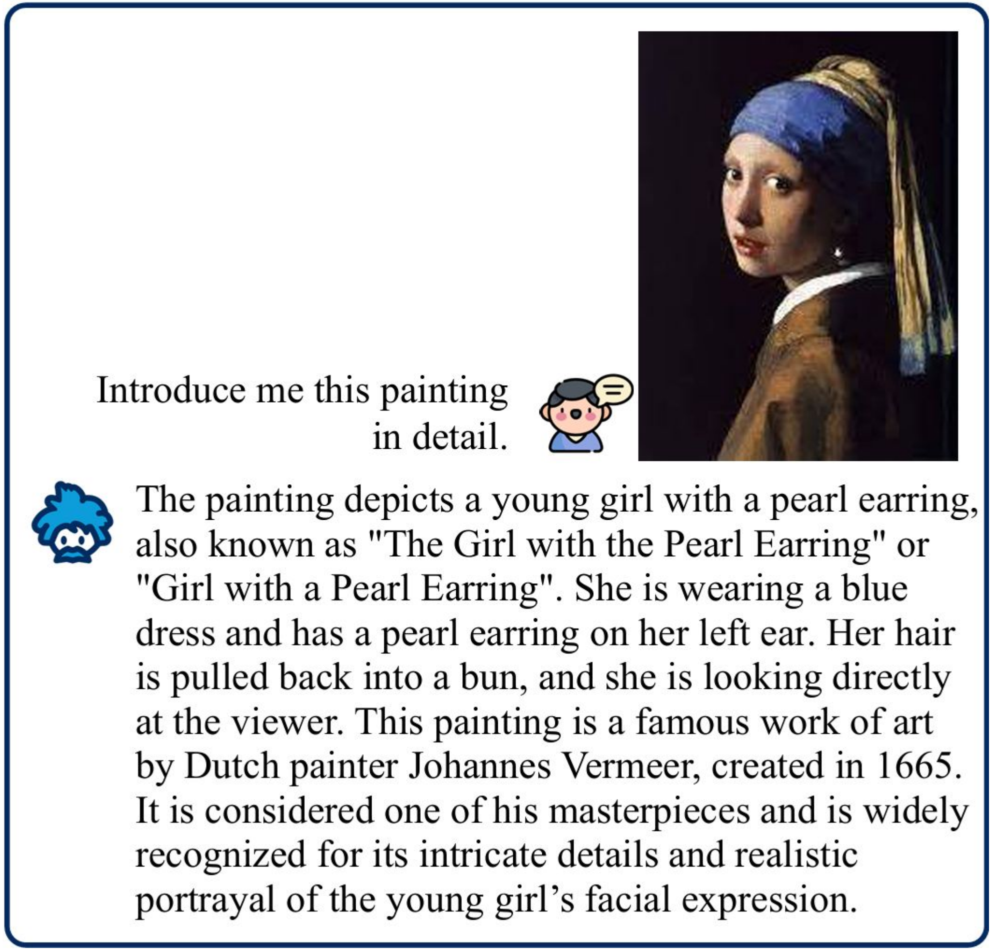
- New zero-shot SOTA results on **all datasets**
- Up to 47.1% relative improvement on MSRVTT-QA over prev. SOTA



What could have happened based on the current scene?

Based on the current scene in the image, it is possible that a hurricane or severe weather event caused significant damage to the buildings and infrastructure in the area. The person walking through the debris-covered area suggests that they are inspecting the damage caused by the disaster. The presence of palm trees indicates that the location may be in a tropical or sub-tropical region, which can be more susceptible to hurricanes and other natural disasters.

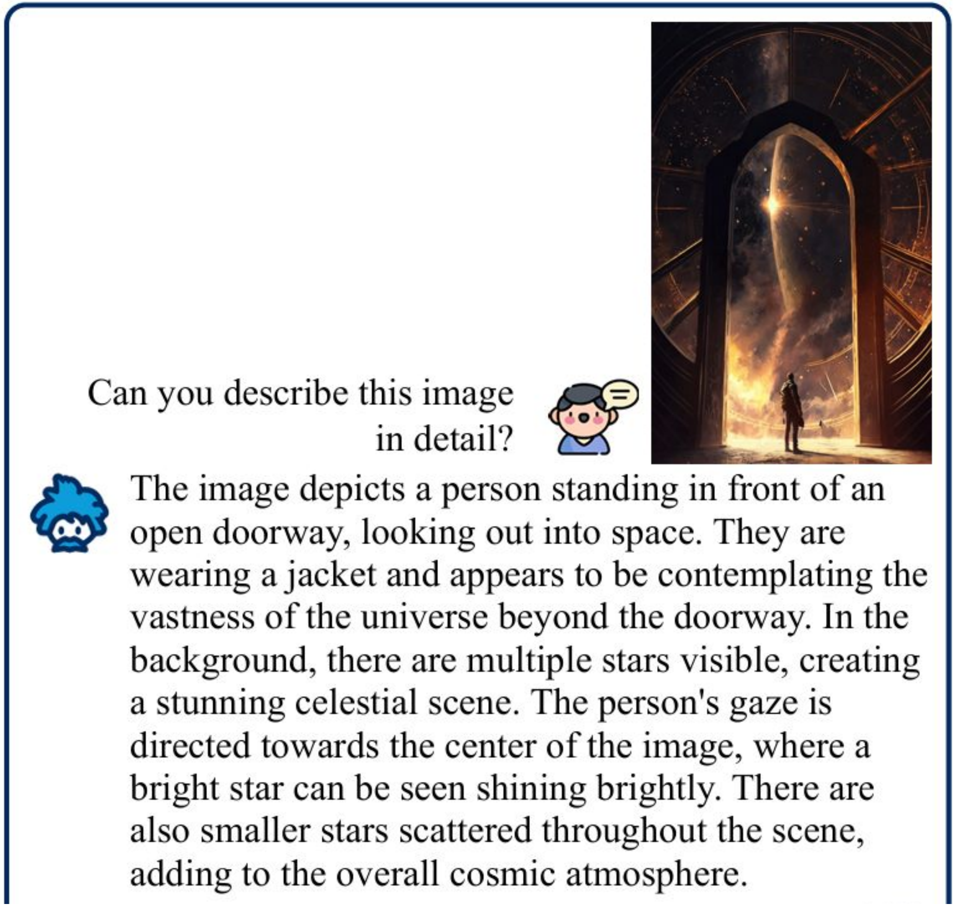
Reasonable inferences → Complex visual reasoning



Introduce me this painting in detail.

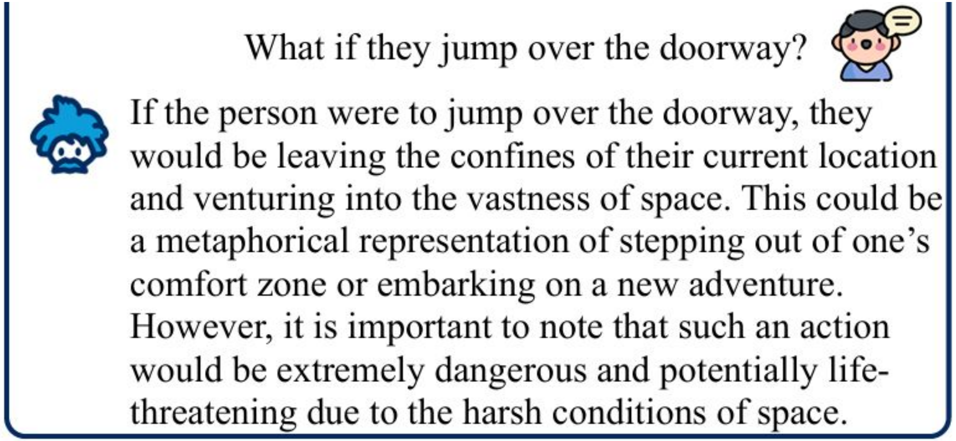
The painting depicts a young girl with a pearl earring, also known as "The Girl with the Pearl Earring" or "Girl with a Pearl Earring". She is wearing a blue dress and has a pearl earring on her left ear. Her hair is pulled back into a bun, and she is looking directly at the viewer. This painting is a famous work of art by Dutch painter Johannes Vermeer, created in 1665. It is considered one of his masterpieces and is widely recognized for its intricate details and realistic portrayal of the young girl’s facial expression.

Connecting visual input with textual knowledge → knowledge-grounded image description



Can you describe this image in detail?

The image depicts a person standing in front of an open doorway, looking out into space. They are wearing a jacket and appears to be contemplating the vastness of the universe beyond the doorway. In the background, there are multiple stars visible, creating a stunning celestial scene. The person's gaze is directed towards the center of the image, where a bright star can be seen shining brightly. There are also smaller stars scattered throughout the scene, adding to the overall cosmic atmosphere.



What if they jump over the doorway?

If the person were to jump over the doorway, they would be leaving the confines of their current location and venturing into the vastness of space. This could be a metaphorical representation of stepping out of one's comfort zone or embarking on a new adventure. However, it is important to note that such an action would be extremely dangerous and potentially life-threatening due to the harsh conditions of space.

- Recognizing metaphorical implications behind image
- Multi-turn conversations

Discussing the use of **off-the-shelf frozen LLMs**

- Model can be adapted to a wide range of LLMs
- **Inherits flaws**, e.g., hallucinations and biases
  - Reduced through
    - diverse datasets
    - grounding responses in vision and instruction input (evidence)
- Cannot finetune LLM parameters → precise following of given instructions not controllable

Takeaway

InstructBLIP enables VLMs to generate responses grounded in what they see and what they're asked.

Impact

- HuggingFace
- MiniGPT-4
- LLaVA Benchmarking

Held-Out Validation Datasets

- NoCaps contains 15,100 images with 166,100 human-written captions for novel object image captioning.
- Visual dialog is a conversational question answering dataset. We use the val split as the held-out,which contains 2,064 images and each has 10 rounds.
- TextVQA requires models to comprehend visual text to answer questions.
- A binary classification dataset to justify whether a meme contains hateful content.

Held-Out Test Datasets

- The Flickr 30k dataset consists of 31K images collected from Flickr, each image has five ground truth captions. We use the test split as the held-out which contains 1K images.
- VizWiz: A dataset contains visual questions asked by people who are blind. 8K images are used for the held-out evaluation.
- GQA contains image questions for scene understanding and reasoning. We use the balanced test-dev set as held-out.
- VSR is a collection of image-text pairs, in which the text describes the spatial relation of two objects in the image. Models are required to classify true/false for the description. We use the zero-shot data split given in its official github repository.

Held-Out Test Datasets (cont.)

- IconQA measures the abstract diagram understanding and comprehensive cognitive reasoning abilities of models. We use the test set of its multi-text-choice task for held-out evaluation.
- ScienceQA covers diverse science topics with corresponding lectures and explanations. In oursettings, we only use the part with image context (IMG).
- MSVD-QA: We use the test set (13K video QA pairs) of MSVD-QA for held-out testing
- MSRVTT-QA has more complex scenes than MSVD, with 72K video QA pairs as the test set.
- iVQA is a video QA dataset with mitigated language biases. It has 6K/2K/2K samples for train/val/test.

Ablation Study on Instruction Tuning

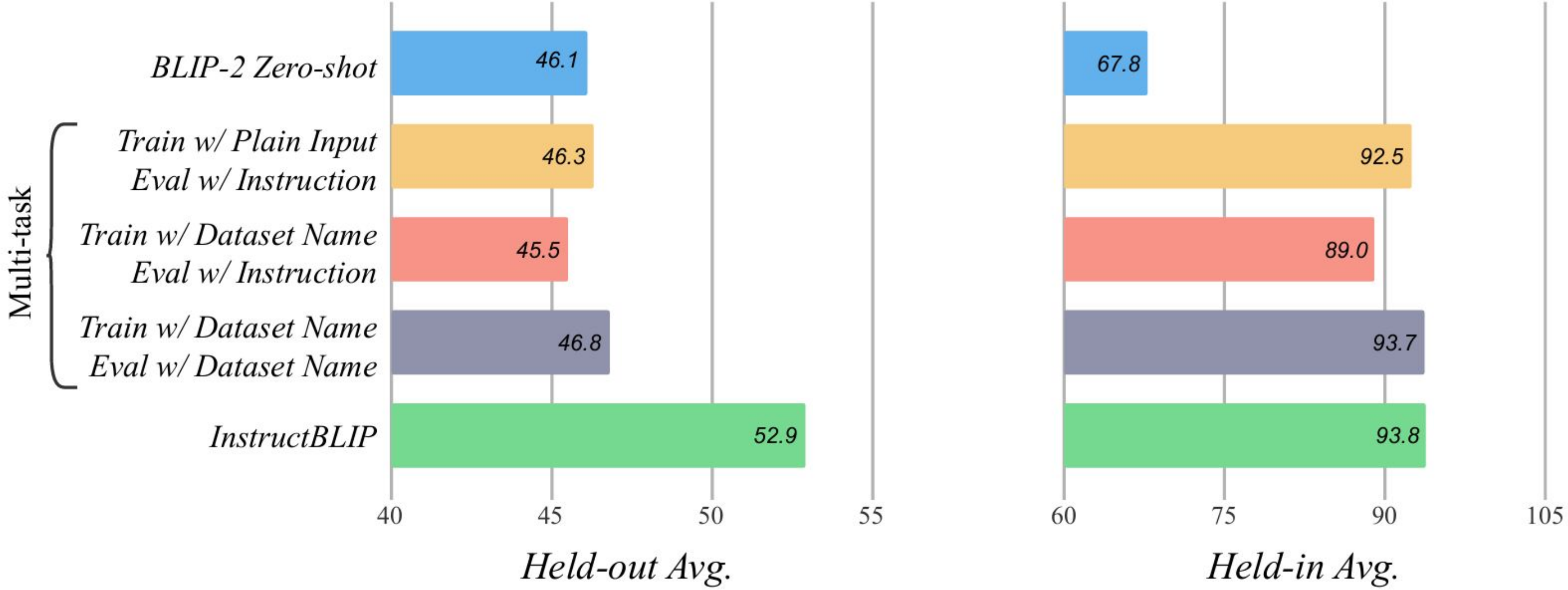
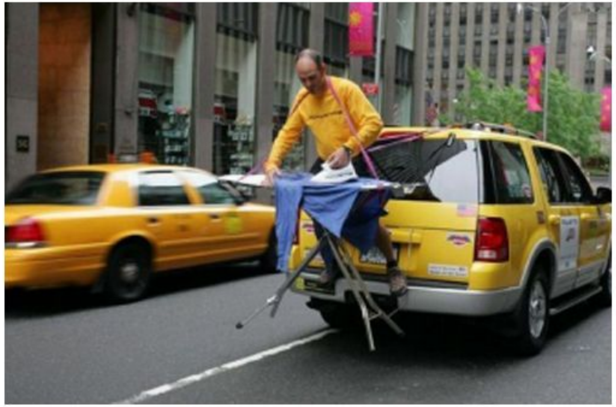


Figure 4: Comparison of instruction tuning and multitask training based on BLIP-2 FlanT5$_{\text{XL}}$ backbone. For held-in evaluation, we compute the average score across all held-in datasets. For held-out evaluation, we compute the average score across GQA, TextVQA, VSR, HatefulMemes, IconQA, ScienceQA, iVQA, VizWiz.

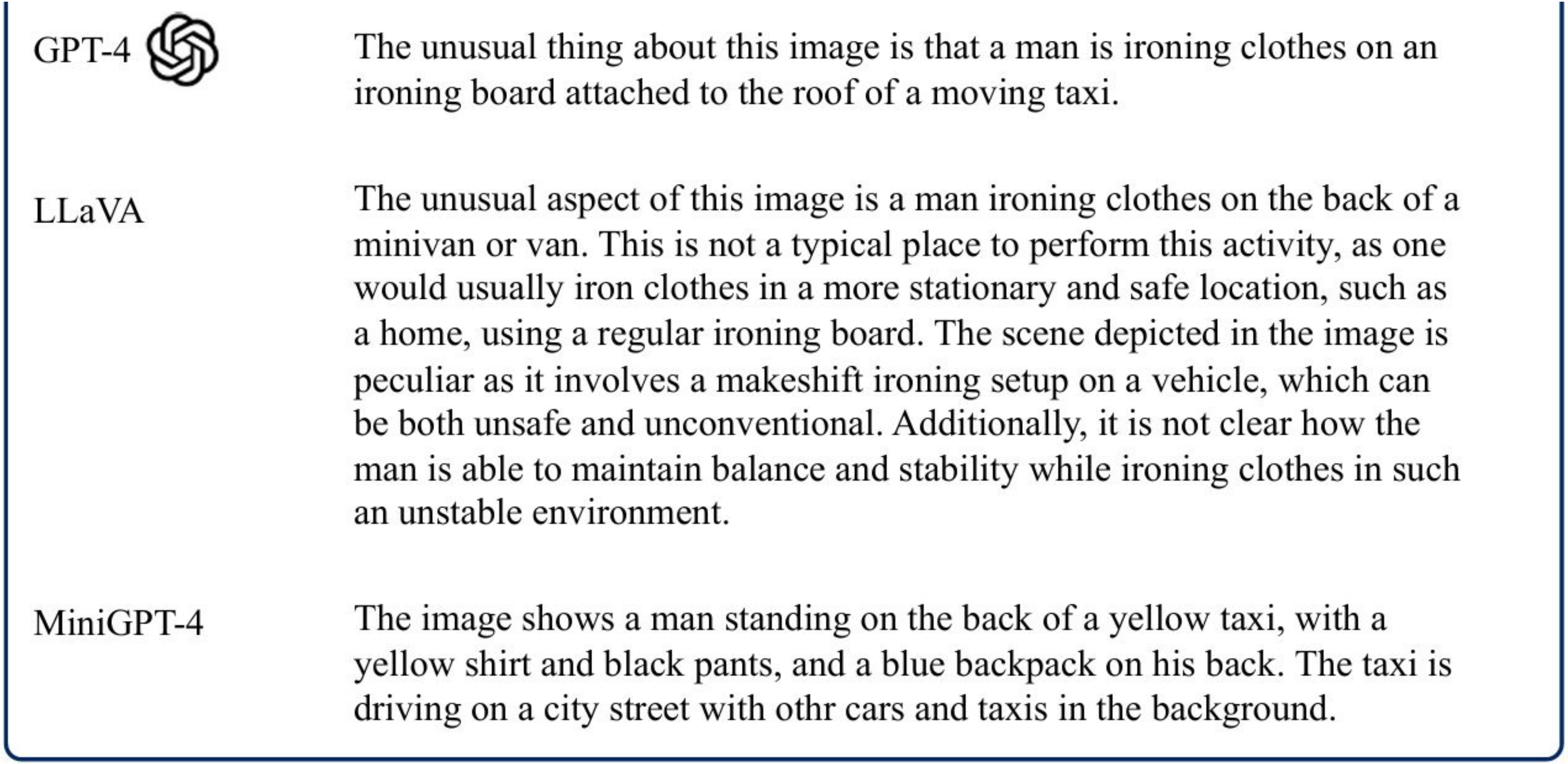
[Input Image]



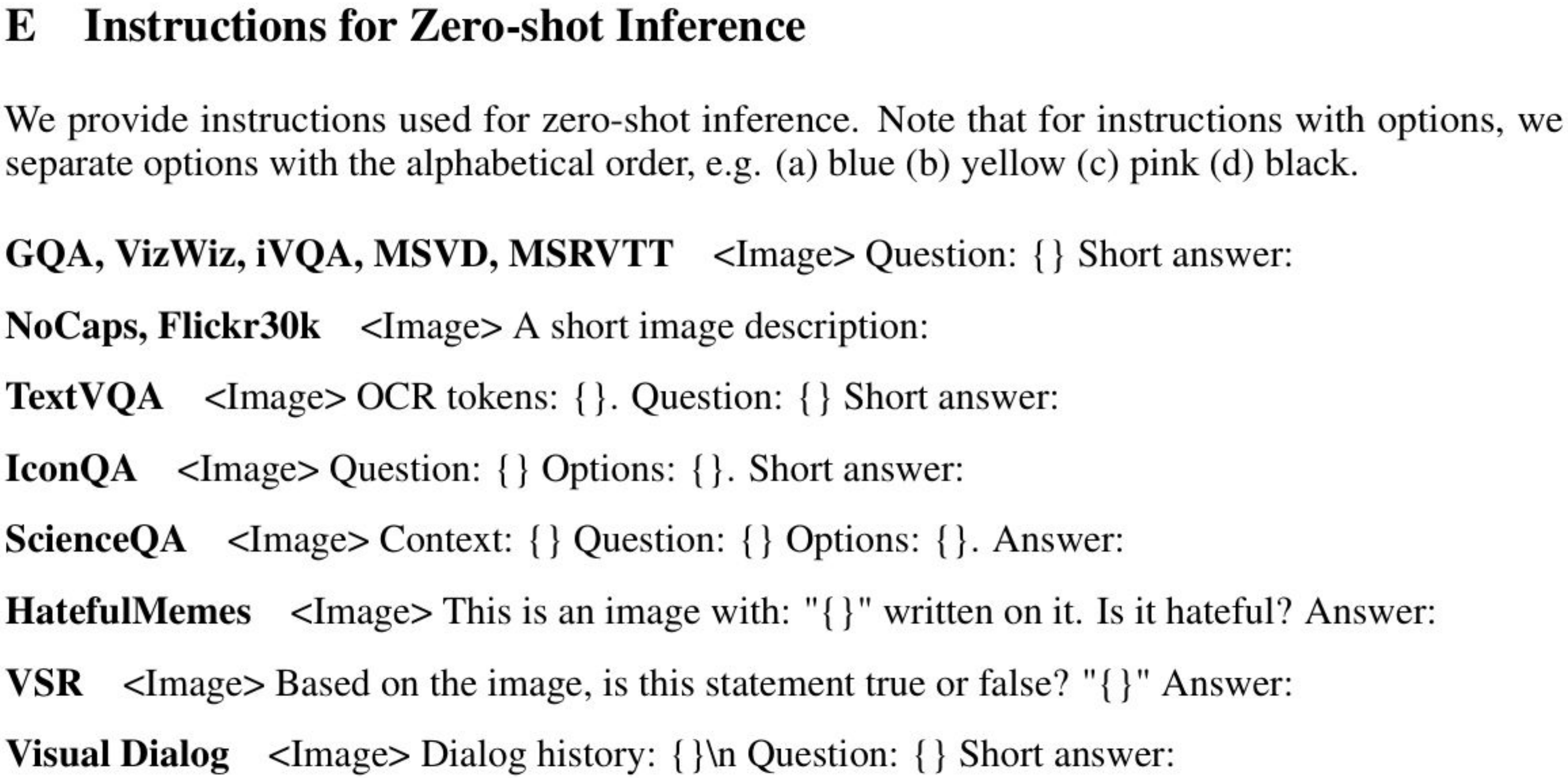
[Input Instruction] What is unusual about this image?

[Model Response]

InstructBLIP The unusual aspect of this image is that a man is ironing clothes on the back of a yellow taxi cab while it is driving down a bus city street. He is standing on top of the car's trunk, holding an iron and using it to press his clothing items. This situation is not only unconventional but also potentially hazardous, as the man could be at risk of falling off the vehicle or causing an accident due to his actions.



GPT-4	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
MiniGPT-4	The image shows a man standing on the back of a yellow taxi, with a yellow shirt and black pants, and a blue backpack on his back. The taxi is driving on a city street with othr cars and taxis in the background.



E Instructions for Zero-shot Inference

We provide instructions used for zero-shot inference. Note that for instructions with options, we separate options with the alphabetical order, e.g. (a) blue (b) yellow (c) pink (d) black.

GQA, VizWiz, iVQA, MSVD, MSRVTT <Image> Question: {} Short answer:

NoCaps, Flickr30k <Image> A short image description:

TextVQA <Image> OCR tokens: {}. Question: {} Short answer:

IconQA <Image> Question: {} Options: {}. Short answer:

ScienceQA <Image> Context: {} Question: {} Options: {}. Answer:

HatefulMemes <Image> This is an image with: "{}" written on it. Is it hateful? Answer:

VSR <Image> Based on the image, is this statement true or false? "{}" Answer:

Visual Dialog <Image> Dialog history: {}\n Question: {} Short answer: