A Broader Impact

InstructBLIP uses off-the-shelf frozen LLMs. Therefore it inherits some of the shortcomings from the original LLMs, such as hallucinating ungrounded text or generating outputs with bias. We mitigate such shortcomings by improving the model's grounding on the vision and instruction input, and performing vision-language instruction tuning on a diverse set of high-quality datasets. Nevertheless, we do not recommend applying InstructBLIP models to any downstream applications without a prior assessment on safety and fairness specific to that application.

B More Case Studies

[Input Image]

[Input Instruction] What is unusual about this image?

[Model Response]

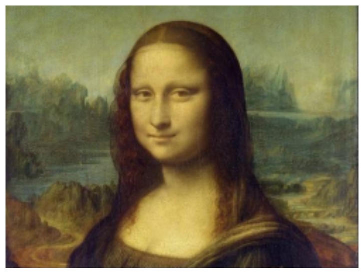
InstructBLIP: The unusual aspect of this image is that a man is ironing clothes on the back of a yellow taxi cab while it is driving down a bus city street. He is standing on top of the car's trunk, holding an iron and using it to press his clothing items. This situation is not only unconventional but 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.

Figure 1: The response from InstructBLIP is more comprehensive than GPT-4, more visually-grounded than LLaVA, and more logical than MiniGPT-4. The responses of GPT-4 and LLaVA are obtained from their respective papers, while the official demo is used for MiniGPT-4. Source: `https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg`

[Input Image] 

[Input Instruction] Do you know who drew this painting?

[Model Response]

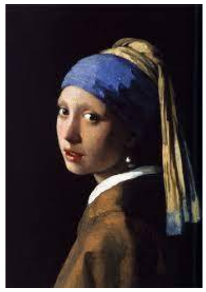
InstructBLIP Leonardo da Vinci.

LLaVA The painting depicts a woman, commonly believed to be Mona Lisa, the famous artwork by Leonardo da Vinci. It is a portrait painting that showcases the woman's enigmatic smile and has become one of the most famous and iconic art pieces in the world. The original work is displayed in the Louvre Museum in Paris, and it is known for its intricate details, use of oil paint, and the artist's innovative techniques that contributed to its enduring appeal and mystery.

MiniGPT-4 The painting is the Mona Lisa, also known as the Mona Lisa, by Leonardo da Vinci. It is one of the most famous paintings in the world and is known for its enigmatic smile and the mysterious identity of the sitter. The painting is housed in the Louvre Museum in Paris, France.

Figure 2: In this example, we illustrate that long-form responses are not always preferable. InstructBLIP can directly address the user's intent by adaptively adjusting the response length, while other models tend to generate lengthy paragraphs with less-relevant sentences. The response from LLaVA is taken from the paper, and for MiniGPT-4, we utilize its official demo.

[Input Image]



[Input Instruction] Introduce me this painting in detail.

[Model Response]

InstructBLIP

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.

LLaVA

The painting features a woman with a headscarf, possibly of Dutch origin, dressed in blue and white clothing. She is wearing a large blue hat and has a red mouth, which adds a vibrant touch to the portrait. The woman is looking directly at the viewer, creating a sense of connection. The background of the painting is predominantly blue, further emphasizing the color theme of the piece. The overall composition and attire of the woman suggest that the painting might be inspired by the style of Rembrandt.

MiniGPT-4

The painting depicts a young woman with a pearl earring, wearing a blue dress and a white apron. She is shown with a serious expression, looking directly at the viewer. The background is a dark, neutral tone, with no visible objects or elements. The overall mood of the painting is somber and contemplative.

The painting is well-executed, with a high level of detail and realism in the depiction of the woman's features and clothing. The use of shading and lighting creates a sense of depth and dimension, making the woman appear three-dimensional. The artist has used a limited color palette, with the woman's dress and apron providing the only sources of color in the painting.

The painting is well-composed, with the woman's head and shoulders positioned in the center of the canvas, creating a sense of balance and symmetry. The use of negative space around the woman's figure adds to the overall somber mood of the painting.

Overall, this is a well-executed and thought-provoking painting that effectively conveys a sense of contemplation and introspection.

Figure 3: InstructBLIP shows the ability to connect textual knowledge with visual input, while the other models depict the picture plainly. MiniGPT-4 exhibits poorer results, which may be due to its training with only long captions. Responses of LLaVA and MiniGPT-4 are generated by their official demos.

C Instruction Tuning Datasets

Dataset Name	Held-out	Dataset Description
COCO Caption [1]	✗	We use the large-scale COCO dataset for the image captioning task. Specifically, Karpathy split [2] is used, which divides the data into 82K/5K/5K images for the train/val/test sets.
Web CapFilt	✗	14M image-text pairs collected from the web with additional BLIP-generated synthetic captions, used in BLIP [3] and BLIP-2 [4].
NoCaps [5]	✓ (val)	NoCaps contains 15,100 images with 166,100 human-written captions for novel object image captioning.
Flickr30K [6]	✓ (test)	The Flickr30k 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.
TextCaps [7]	✗	TextCaps is an image captioning dataset that requires the model to comprehend and reason the text in images. Its train/val/test sets contain 21K/3K/3K images, respectively.
VQAv2 [8]	✗	VQAv2 is dataset for open-ended image question answering. It is split into 82K/40K/81K for train/val/test.
VizWiz [9]	✓ (test-dev)	A dataset contains visual questions asked by people who are blind. 8K images are used for the held-out evaluation.
GQA [10]	✓ (test-dev)	GQA contains image questions for scene understanding and reasoning. We use the balanced test-dev set as held-out.
Visual Spatial Reasoning	✓ (test)	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.
IconQA [11]	✓ (test)	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.
OKVQA [12]	✗	OKVQA contains visual questions that require outside knowledge to answer. It has been split into 9K/5K for train and test.
A-OKVQA [13]	✗	A-OKVQA is a successor of OKVQA with more challenging and diverse questions. It has 17K/1K/6K questions for train/val/test.
ScienceQA [14]	✓ (test)	ScienceQA covers diverse science topics with corresponding lectures and explanations. In out settings, we only use the part with image context (IMG).
Visual Dialog [15]	✓ (val)	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.
OCR-VQA [16]	✗	OCR-VQA contains visual questions that require models to read text in the image. It has 800K/100K/100K for train/val/test, respectively.
TextVQA [17]	✓ (val)	TextVQA requires models to comprehend visual text to answer questions.
HatefulMemes [18]	✓ (val)	A binary classification dataset to justify whether a meme contains hateful content.
LLaVA-Instruct-150K [19]	✗	An instruction tuning dataset which has three parts: detailed caption (23K), reasoning (77K), conversation (58K).
MSVD-QA [20]	✓ (test)	We use the test set (13K video QA pairs) of MSVD-QA for held-out testing.
MSRVTT-QA [20]	✓ (test)	MSRVTT-QA has more complex scenes than MSVD, with 72K video QA pairs as the test set.
iVQA [21]	✓ (test)	iVQA is a video QA dataset with mitigated language biases. It has 6K/2K/2K samples for train/val/test.

Table 1: Description of datasets in our held-in instruction tuning and held-out zero-shot evaluations.

D Instruction Templates

Task	Instruction Template
Image Captioning	<Image>A short image caption:
	<Image>A short image description:
	<Image>A photo of
	<Image>An image that shows
	<Image>Write a short description for the image.
	<Image>Write a description for the photo.
	<Image>Provide a description of what is presented in the photo.
	<Image>Briefly describe the content of the image.
	<Image>Can you briefly explain what you see in the image?
	<Image>Could you use a few words to describe what you perceive in the photo?
	<Image>Please provide a short depiction of the picture.
	<Image>Using language, provide a short account of the image.
	<Image>Use a few words to illustrate what is happening in the picture.
VQA	<Image>{Question}
	<Image>Question: {Question}
	<Image>{Question} A short answer to the question is
	<Image>Q: {Question} A:
	<Image>Question: {Question} Short answer:
	<Image>Given the image, answer the following question with no more than three words. {Question}
	<Image>Based on the image, respond to this question with a short answer: {Question}. Answer:
	<Image>Use the provided image to answer the question: {Question} Provide your answer as short as possible:
	<Image>What is the answer to the following question? "{Question}"
	<Image>The question "{Question}" can be answered using the image. A short answer is
VQG	<Image>Given the image, generate a question whose answer is: {Answer}. Question:
	<Image>Based on the image, provide a question with the answer: {Answer}. Question:
	<Image>Given the visual representation, create a question for which the answer is "{Answer}".
	<Image>From the image provided, craft a question that leads to the reply: {Answer}. Question:
	<Image>Considering the picture, come up with a question where the answer is: {Answer}.
	<Image>Taking the image into account, generate an question that has the answer: {Answer}. Question:

Table 2: Instruction templates used for transforming held-in datasets into instruction tuning data. For datasets with OCR tokens, we simply add "OCR tokens:" after the image query embeddings.

References

[1] 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 *ECCV*, 2014.

[2] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[3] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[5] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, pages 8948–8957, 2019.

[6] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2, 2014.

[7] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioningwith reading comprehension. 2020.

[8] 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 *CVPR*, July 2017.

[9] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, 2018.

[10] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

[11] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS Track on Datasets and Benchmarks*, 2021.

[12] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

[13] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, 2022.

[14] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

[15] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017.

[16] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

[17] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019.

[18] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020.

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023.

[20] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 1645–1653, 2017.

[21] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1686–1697, 2021.