# Vision-Flan: Scaling Human-Labeled Tasks in Visual Instruction Tuning

**Zhiyang Xu** **Ying Shen** **Trevor Ashby** **Lifu Huang**
Department of Computer Science, Virginia Tech
{zhiyangx, yings, trevorashby, lifuh}@vt.edu

## Abstract

Despite vision-language models' (VLMs) remarkable capabilities as versatile visual assistants, two substantial challenges persist within the existing VLM frameworks: (1) *lacking task diversity* in pretraining and visual instruction tuning, and (2) *annotation error* and *bias* in GPT-4 synthesized instruction tuning data. Both challenges lead to issues such as poor generalizability, hallucination, and catastrophic forgetting. To address these challenges, we propose VISION-FLAN, the most diverse public-available visual instruction tuning dataset to date, comprising 196 diverse tasks and 1,664,261 instances sourced from academic datasets, and each task is accompanied by an expert-written instruction. Complementing the proposed dataset, we further introduce a two-stage instruction tuning framework, in which VLMs are firstly tuned on VISION-FLAN and secondly, further tuned on GPT-4 synthesized data. Our experimental results demonstrate that by leveraging the two-stage tuning framework, VLMs trained on VISION-FLAN, achieve the state-of-the-art performance across a wide range of multi-modal evaluation benchmarks.

## 1 Introduction

Recent vision-language models (VLMs) (Liu et al., 2023d; Li et al., 2023b; Dai et al., 2023), built upon pre-trained large-language models (LLMs) (Chiang et al., 2023; Gao et al., 2023) and pretrained image encoders (Sun et al., 2023), have shown impressive capabilities as general visual assistants. However, despite their notable successes, we identify two remaining challenges that merit further investigation.

**Firstly**, the data used in the pre-training stage is dominated by the image captioning task, which lacks diversity, resulting in limited generalizability of VLMs (Chen et al., 2023; Zhang et al., 2023). **Secondly**, most of existing visual instruction tuning datasets (Liu et al., 2023d; Li et al., 2023a; Yin et al., 2023) are synthetically generated by GPT-4 by repurposing text annotations from the original computer-vision datasets. The lack of task diversity, spurious co-occurring patterns between objects, and long-form outputs in these datasets may cause severe hallucination (Liu et al., 2023b; Li et al., 2023c; Liu et al., 2023a; Zhou et al., 2023), and catastrophic forgetting (Zhai et al., 2023).

To address both challenges, we introduce VISION-FLAN, the most diverse public-available visual instruction tuning dataset consisting of 196 tasks drawn from academic datasets. Each task in VISION-FLAN is accompanied by an expert-written instruction. We show some sample tasks from VISION-FLAN in Figure 2 and all the datasets used in Appendix B. In addition, we introduce a novel two-stage instruction tuning framework. In the first stage, we utilize the pre-trained LLaVA model (Liu et al., 2023d) as our initial model, and finetune it on VISION-FLAN to gain diverse capabilities, resulting in the VISION-FLAN BASE model. However, due to the concise nature of target outputs in academic datasets, the responses generated by VISION-FLAN BASE tend to be brief and not aligned with human preferences. Therefore, in the second stage, we further finetune VISION-FLAN BASE using a minimal amount of GPT-4 synthesized data (i.e., 1,000). This step aims to adjust the model's outputs to be more in line with human preference, resulting in the VISION-FLAN CHAT model.

Our experimental results demonstrate that high-quality human annotations within VISION-FLAN significantly enhances the capabilities of both VISION-FLAN BASE and VISION-FLAN CHAT while reducing the risk of hallucination and catastrophic forgetting. The two-stage instruction tuning framework enables VISION-FLAN CHAT to achieve better human-preference alignment with much less GPT-4 synthesized data comparing to state-of-the-art VLMs.

| Model | LLM | Image Encoder | MM-Bench | MME | LLaVA-Bench | MM-Vet | Pope | CF |
|---|---|---|---|---|---|---|---|---|
| BLIP-2 | FlanT5-XXL | ViT-g/14 | - | 1293.8 | - | 22.4 | 85.3 | - |
| InstructBlip | Vicuna-13B | ViT-g/14 | 36.0 | 1212.8 | 58.2 | 25.6 | 78.9 | - |
| Mini-GPT4 | Vicuna-13B | ViT-g/14 | 24.3 | 581.67 | - | - | - | - |
| Shikra | Vicuna-13B | ViT-L/14 | 58.8 | - | - | - | - | - |
| LLaVA | Vicuna-13B v1.5 | CLIP-ViT-L-336px | 38.7 | 1151.6 | 70.8 | 33.4 | 75.3 | - |
| Qwen-VL | Qwen-7B | ViT-bigG | 38.2 | - | - | - | - | - |
| Qwen-VL-Chat | Qwen-7B | ViT-bigG | 60.6 | 1487.5 | 73.6 | - | - | 72.1 |
| LLaVA 1.5 | Vicuna-13B v1.5 | CLIP-ViT-L-336px | 66.7 | 1531.3 | 70.7 | 35.4 | 83.6 | 73.3 |
| VISION-FLAN BASE | Vicuna-13B v1.5 | CLIP-ViT-L-336px | **69.8** | **1537.8** | 38.5 | 33.4 | 85.9 | **87.2** |
| **Second-Stage Alignment with 1,000 LLaVA** | | | | | | | | |
| VISION-FLAN CHAT | Vicuna-13B v1.5 | CLIP-ViT-L-336px | 67.6 | 1490.6 | **78.3** | **38.0** | **86.1** | 84.0 |

Table 1: Comprehensive evaluation of VLMs on widely adopted benchmark datasets.

## 2 Two-stage Visual Instruction Tuning

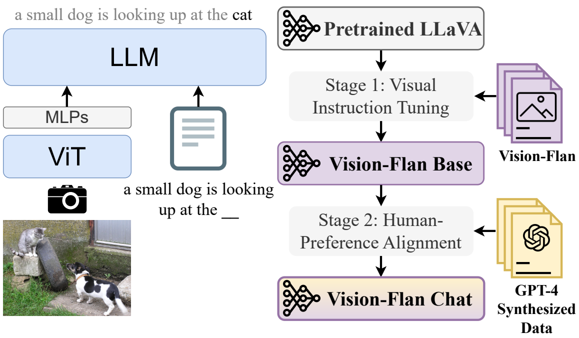


Figure 1: On the left of the figure, we show the architecture of the LLaVA model and on the right of the figure, we show the pipeline of the two-stage visual instruction tuning.

Contrary to prior approaches (Liu et al., 2023c; Dai et al., 2023) that mix human-labeled data with GPT-4 synthesized data for visual instruction tuning, our study introduces a two-stage instruction tuning pipeline. In the first stage, we finetune the VLM on VISION-FLAN to acquire diverse capabilities and name the resulting model as VISION-FLAN BASE. However, due to the brevity of target outputs presenting in the academic datasets, the responses from VISION-FLAN BASE are not in human-preferred formats. Hence, we further finetune the VLM on GPT-4 synthesized data to align the model's outputs with human preference. We denote the yielded model as VISION-FLAN CHAT.

## 3 Experiment

**Experiment Setup** We evaluate the models on *multiple-choice* benchmarks: **MMbench** (Liu et al., 2023e), and **MME** (Fu et al., 2023); *free-form generation* benchmarks: **MM-Vet** (Yu et al., 2023) and **LLaVA-Bench**; the *hallucination* benchmark: **POPE** (Li et al., 2023c), and *catastrophic forgetting* benchmarks: **CIFAR-10 and CIFAR-100** (Krizhevsky et al., 2009), **MNIST** (LeCun, 1998), and **miniImagenet** (Vinyals et al., 2016).

**Main Results** As demonstrated in Table 1, VISION-FLAN BASE achieves state-of-the-art performance on comprehensive evaluation benchmarks, while reducing hallucination and catastrophic forgetting. However, we observe VISION-FLAN BASE scores significantly lower on the LLaVA-Bench dataset comparing to VLMs trained on GPT-4 synthesized data. We attribute this problem to the conciseness and brevity of target outputs in academic datasets. On the other hand, with the second-stage tuning on a merely 1,000 GPT-4 synthesized instances, VISION-FLAN CHAT achieves significantly improved performance on benchmarks measuring human-preference alignment including LLaVA-Bench and MM-Vet, while maintaining a relatively lower rate of hallucination and catastrophic forgetting.

In Table 3 and 4, we show the effects of using different amount of GPT-4 synthesised data on human-preference alignment and hallucination. As one can observe, A minimal quantity (1,000) of GPT-4 synthesized data is sufficient for aligning VLM responses with human preference. Notably, an increase in the number of GPT-4 synthesized data does not correspond to a proportional enhancement in alignment and introduces hallucination and bias into the VLMs.

## 4 Conclusion

In this paper, we propose VISION-FLAN, the most diverse public-available visual instruction tuning dataset, consisting of 196 diverse tasks and 1,664,261 instances collected from academic datasets, and and each task is accompanied by an expert-written instruction. We demonstrate that VLMs trained on VISION-FLAN with the proposed two-stage visual instruction tuning framework achieve state-of-the-art performance on comprehensive evaluation bemchmark datasets.

## References


Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9448–9458. Curran Associates, Inc.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *CoRR*, abs/2302.11713.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. July 8-10, 2009. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece.

Chee Kheng Ch'ng, Chee Seng Chan, and Chenglin Liu. 2020. Total-text: Towards orientation robustness in scene text detection. *International Journal on Document Analysis and Recognition (IJDAR)*, 23:31–52.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Adam Coates, Andrew Y. Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, pages 215–223. JMLR.org.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. 2018. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):44:1–44:10.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1778–1785. IEEE.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Noa Garcia and George Vogiatzis. 2018. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2019. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

Gregory Griffin, Alex Holub, and Pietro Perona. 2007. Caltech-256 object category dataset.

Jean-Philippe Thiran Guillaume Jaume, Hazim Kemal Ekenel. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *Accepted to ICDAR-OST*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII*, volume 12362 of *Lecture Notes in Computer Science*, pages 417–434. Springer.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural adversarial examples. *CVPR*.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

EunJeong Hwang and Vered Shwartz. 2023. Memecap: A dataset for captioning and interpreting memes.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017a. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017b. Inferring and executing programs for visual reasoning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3008–3017. IEEE Computer Society.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *CVPR*.

Yannis Kalantidis, Lluis Garcia Pueyo, Michele Trevisiol, Roelof van Zwol, and Yannis Avrithis. 2011. Scalable triangulation-based logo recognition. In *Proceedings of the 1st ACM international conference on multimedia retrieval*, pages 1–7.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561.

Elisa Kreiss, Fei Fang, Noah D. Goodman, and Christopher Potts. 2022. Concadia: Towards image-based text generation with a purpose.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Anurendra Kumar, Keval Morabia, William Wang, Kevin Chang, and Alex Schwing. 2022. CoVA: Context-aware visual attention for webpage information extraction. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 80–90, Dublin, Ireland. Association for Computational Linguistics.

Jason J Lau, Soumya Gayen, Dina Demner, and Asma Ben Abacha. 2019. Visual question answering in radiology (vqa-rad).

Yann LeCun. 1998. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G Moreno, and Jesús Lovón Melgarejo. 2022. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'22, New York, NY, USA. Association for Computing Machinery.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. 2017. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5543–5551. IEEE Computer Society.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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

Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. 2019. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2152–2161.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023e. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuen Peng Loh and Chee Seng Chan. 2019. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42.

Vincenzo Lomonaco and Davide Maltoni. 2017. Core50: a new dataset and benchmark for continuous object recognition. In *Conference on robot learning*, pages 17–26. PMLR.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. 2013. Fine-grained visual classification of aircraft. Technical report.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for VQA on document images. *CoRR*, abs/2007.00398.

Alexander Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model.

Alex Olsen, Dmitry A. Konovalov, Bronson Philippa, Peter Ridd, Jake C. Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, Brendan Calvert, Mostafa Rahimi Azghadi, and Ronald D. White. 2019. DeepWeeds: A Multiclass Weed Species Image Dataset for Deep Learning. *Scientific Reports*, 9(2058).

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. 2017a. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. 2017b. Visda: The visual domain adaptation challenge.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *ECCV*.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, volume 6314 of *Lecture Notes in Computer Science*, pages 213–226. Springer.

Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2016. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18.

Naganand Yadati Sanket Shah, Anand Mishra and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *AAAI*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

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

Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. 2019. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1221–1231.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. 2019. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. 2017. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027.

Manisha Verma, Sudhakar Kumawat, Yuta Nakashima, and Shanmuganathan Raman. 2020. Yoga-82: A new dataset for fine-grained classification of human poses. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4472–4479.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. Ucsd birds. Technical Report CNS-TR-2011-001, California Institute of Technology.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. 2017. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. 2023. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Liu Yuliang, Jin Lianwen, Zhang Shuaitao, and Zhang Sheng. 2017. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. 2019. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models.

Yutong Zhou and Nobutaka Shimada. 2021. Generative adversarial network for text-to-face synthesis and manipulation with pretrained bert model. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08.

# A Sample Tasks

# B Datasets Used in VISION-FLAN

CINIC-10 (Darlow et al., 2018), MSCOCO (Lin et al., 2014), FairFace (Karkkainen and Joo, 2021), IconQA (Lu et al., 2021b), ImageNet-A (Hendrycks et al., 2021b), ImageNet-C (Hendrycks and Dietterich, 2019), InfographicVQA (Mathew et al., 2022), SemArt (Garcia and Vogiatzis, 2018) (Bevilacqua et al., 2012), TextCaps (Sidorov et al., 2020), VisDial (Das et al., 2017), VizWiz (Gurari et al., 2018), STL-10 (Coates et al., 2011), Office-31 (Saenko et al., 2010), LSUN (Yu et al., 2015), FGVC-Aircraft (Maji et al., 2013), DeepFashion (Liu et al., 2016), CUB-200-2011 (Wah et al., 2011), CLEVR (Johnson et al., 2017a), A-OKVQA (Schwenk et al., 2022), AI2D (Kembhavi et al., 2016), AID (Xia et al., 2017), Caltech-256 (Griffin et al., 2007), CoVA (Kumar et al., 2022), DeepWeeds (Olsen et al., 2019), ExDark (Loh and Chan, 2019), FFHQ-Text (Zhou and Shimada, 2021), FlickrLogos-27 (Kalantidis et al., 2011), ImageNet-R (Hendrycks et al., 2021a), ImageNet-Sketch (Wang et al., 2019), JHU-CROWD++ (Sindagi et al., 2019), MNIST-M (Ganin et al., 2016), MVTecAD (Bergmann et al., 2019), NABirds (Van Horn et al., 2015), Road-Anomaly (Lis et al., 2019), SCUT-CTW1500 (Yuliang et al., 2017),

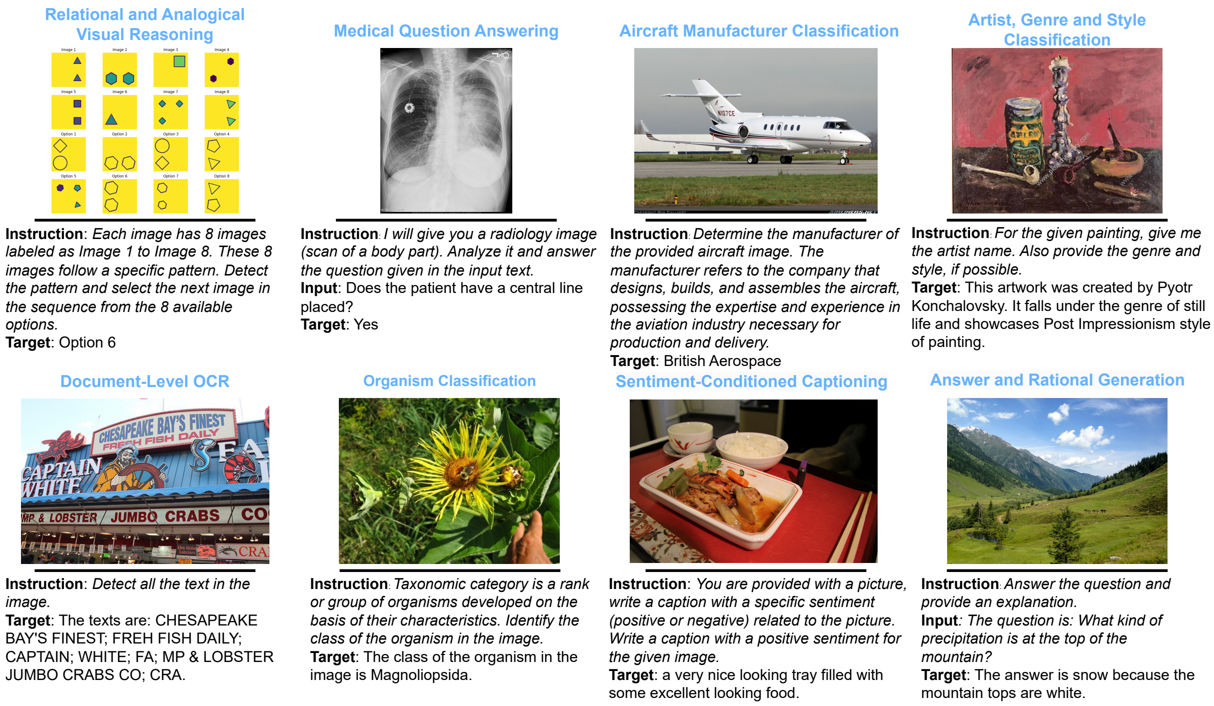


Figure 2: Sample tasks in VISION-FLAN. **Instruction** denotes a task instruction crafted by annotators. **Input** means text input in the given task, and **Target** is the target response based on the instruction.

Total-Text (Ch'ng et al., 2020), VisDA-2017 (Peng et al., 2017a), Yoga-82 (Verma et al., 2020), Caltech101 (Fei-Fei et al., 2004), Cars (Krause et al., 2013), Core50 (Lomonaco and Maltoni, 2017), NUS-WIDE (Chua et al., July 8-10, 2009), ObjectNet (Barbu et al., 2019), Places205 (Zhou et al., 2014), 300w (Sagonas et al., 2016), Yahoo (Farhadi et al., 2009), LFW (Huang et al., 2007), model-vs-human (Geirhos et al., 2019), Office-Home (Venkateswara et al., 2017), Winoground (Thrush et al., 2022), ConceptualCaptions (Sharma et al., 2018), KVQA+image question answer (Sanket Shah and Talukdar, 2019), MemeCap (Hwang and Shwartz, 2023), PlotQA (Methani et al., 2020), SentiCap (Mathews et al., 2016), VisDA-2017 (Peng et al., 2017b), VQG (Mostafazadeh et al., 2016), WIT (Srinivasan et al., 2021), WikiArt (Tan et al., 2019), VQA-RAD (Lau et al., 2019), VOC2007 (Everingham et al.), VIZWIZ (Gurari et al., 2020), ViQuAE (Lerner et al., 2022), ST-VQA (Biten et al., 2019), Sketch (Eitz et al., 2012), RAVEN (Zhang et al., 2019), PICKAPIC (Kirstain et al., 2023), PACS (Li et al., 2017), NOCAPS (Agrawal et al., 2019), Localized Narratives (Pont-Tuset et al., 2020), INATURALIST (Horn et al., 2018), HICO (Chao et al., 2015), GEOMETRY3K (Lu et al., 2021a), FUNSD (Guillaume Jaume, 2019), FLICKR30K (Plummer et al., 2017), DVQA (Kafle et al., 2018), DTD (Cimpoi et al., 2014), DOMAIN NET (Peng et al., 2019), DOCVQA (Mathew et al., 2020), DAQUAR (Malinowski and Fritz, 2014), CONCADIA (Kreiss et al., 2022), CLEVR (Johnson et al., 2017b), and CHART2TEXT (Obeid and Hoque, 2020).

## C Effect of GPT-4 Synthesized Data

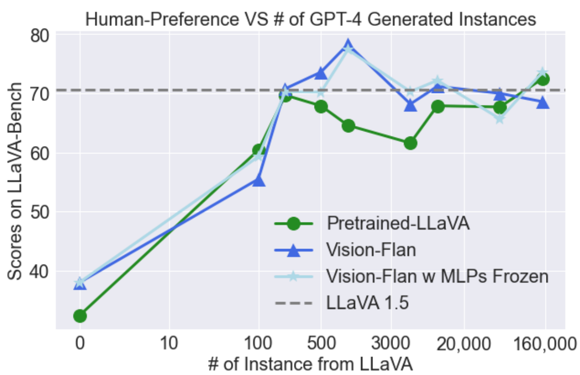


Figure 3: Effect of increasing number of GPT-4 synthesized training instances on the human-preference benchmark. The dashed gray line indicates the performance of the-state-of-the-art LLaVA 1.5 model.

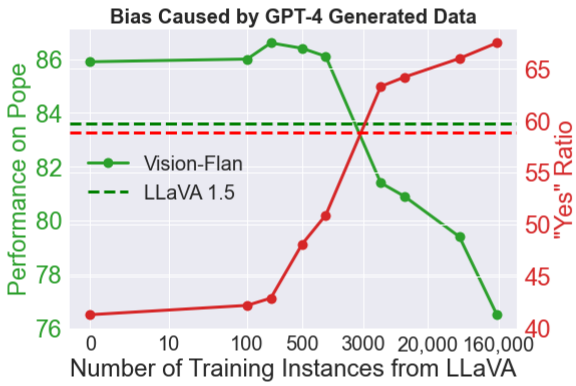


Figure 4: Effect of increasing number of GPT-4 synthesized training instances on the hallucination benchmark and the ratio of “Yes”. The dashed lines indicate the performance of the state-of-the-art LLaVA 1.5 model.