HOW WELL DOES GPT-4V(ISION) ADAPT TO DISTRIBUTION SHIFTS? A PRELIMINARY INVESTIGATION

Zhongyi Han[1*], **Guanglin Zhou**[2*], **Rundong He**[3*], **Jindong Wang**[4], **Tailin Wu**[5], **Yilong Yin**[3], **Salman Khan**[1,6], **Lina Yao**[7,2,9], **Tongliang Liu**[9,1] & **Kun Zhang**[10,1]
[1]Mohamed bin Zayed University of Artificial Intelligence [2]The University of New South Wales
[3]School of Software, Shandong University [4]Microsoft Research Asia [5]Westlake University
[6]Australian National University [7]Data61, CSIRO [8]Macquarie University
[9]The University of Sydney [10]Carnegie Mellon University
{hanzhongyicn;jameszhou.ustc}@gmail.com

ABSTRACT

In machine learning, generalization against distribution shifts is crucial, particularly in fields like climate modeling, biomedicine, and autonomous driving. The emergence of foundation models has led to an increased interest in their adaptability to distribution shifts. GPT-4V(ision) acts as one of the most advanced publicly accessible multimodal foundation models, with extensive applications across various domains. However, its robustness against data distributions remains largely underexplored. Addressing this gap, this study rigorously evaluates GPT-4V's adaptability and generalization capabilities in dynamic environments, benchmarking against prominent models like CLIP, LLaVA, and Gemini. We delve into GPT-4V's zero-shot generalization across 13 diverse datasets spanning natural, medical, and molecular domains. We further investigate its adaptability to controlled data perturbations and examine the efficacy of in-context learning as a tool to enhance its adaptation. Our findings delineate GPT-4V's capability boundaries in distribution shifts, shedding light on its strengths and limitations across various scenarios. Importantly, this investigation contributes to our understanding of how AI foundation models generalize to distribution shifts, offering pivotal insights into their adaptability and robustness. Code is publicly available at https://github.com/jameszhou-ql/gpt-4v-distribution-shift.

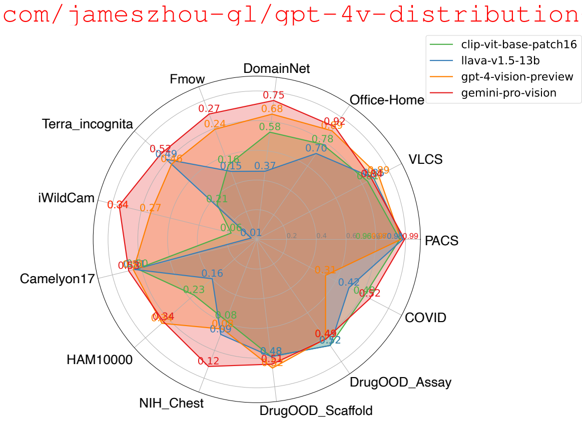


Figure 1: Comparative analysis of zero-shot generalization performance across 13 distinct datasets, encompassing natural, medical, and molecular domains. The analysis features the performances of three advanced models: GPT-4V, CLIP, LLaVA, and Gemini.

1 INTRODUCTION

In the dynamic field of machine learning, the phenomenon of distribution shift poses a significant challenge to the generalization capabilities of algorithms. The divergence between training data

*These authors contributed equally to this work.

distributions and those encountered in real-world scenarios can severely impact model performance across various domains such as climate modeling, biomedicine, wildlife conservation, autonomous driving, and financial forecasting (Knutti et al., 2010; Park et al., 2021; Tuia et al., 2022; Stocco & Tonella, 2022; Mashrur et al., 2020). This issue highlights the limitations of the traditional independent and identically distributed (i.i.d.) assumption, underscoring the necessity for models that can adapt and maintain accuracy amidst evolving data distributions.

Seminal work in domain adaptation has demonstrated progress in enhancing model robustness by addressing known distribution shifts during training (Sun & Saenko, 2016; Ganin & Lempitsky, 2015). Yet, the advent of foundation models, exemplified by GPT-4V(ision) (hereafter "GPT-4V"), marks a pivotal advancement in confronting these shifts through extensive pretraining and large-scale data comprehension (Bommasani et al., 2021; Radford et al., 2021). Despite the broad applications and theoretical adaptability of GPT-4V, its practical resilience to distribution shifts remains underexplored, particularly in high-stakes scenarios where reliability is paramount.

This study embarks on a new exploration of GPT-4V's ability to distribution shifts, setting a benchmark against models renowned for their robustness, such as CLIP, LLaVA, and Gemini (Radford et al., 2021; Liu et al., 2023b; Team et al., 2023). We dissect GPT-4V's performance across a spectrum of distribution shift scenarios. Our investigation is structured around three critical questions:

1. *How effectively does GPT-4V manage distribution shifts across diverse domains?* We evaluate GPT-4V's zero-shot adaptability to various domain-specific distribution shifts, benchmarking against models with established robustness.
2. *How does GPT-4V react to deliberate alterations in data distribution?* This examines GPT-4V's generalization capabilities under controlled distribution shifts, including Gaussian noise and stylistic transformations.
3. *Is in-context learning an effective method to augment GPT-4V's adaptation to distribution shifts?* We explore the potential of in-context learning as an alternative to traditional domain generalization methods, assessing its impact on GPT-4V's adaptability.

2 Results

2.1 Zero-shot Recognition

In the realm of zero-shot generalization across various natural, medical, and scientific image datasets, GPT-4V (gpt-4-vision-preview) demonstrates impressive capabilities, often surpassing the performance of baseline models such as CLIP (clip-vit-base-patch16) and LLaVA (llava-v1.5-13b). However, GPT-4V underperforms Gemini (gemini-pro-vision) in most scenarios.

Table 1: Summary of zero-shot generalization performance across various natural datasets, showcasing the comparative results of GPT-4V (gpt-4-vision-preview) with CLIP (clip-vit-base-patch16), LLaVA (llava-v1.5-13b), and Gemini (gemini-pro-vision) models.

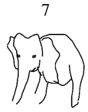
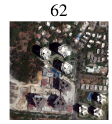
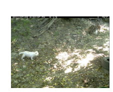
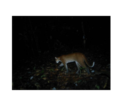
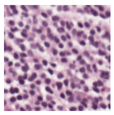
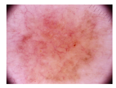
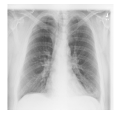
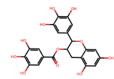
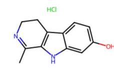
Dataset	PACS	VLCS	Office-Home	DomainNet	Fmow	TerraIncognita	iWildCam
Category	natural	natural	natural	natural	natural	natural	natural
Prediction	animal species	animal species	everyday items	objects, creatures	land use	animal species	animal species
Domain	artistic media	image repositories	visual categories	artistic styles	time, region	camera trap	location
#domains	4	4	4	6	6	4	206
#classes	7	5	65	345	62	10	323
Examples							
CLIP	0.961 1730/1800	0.808 1455/1800	0.778 1400/1800	0.582 1048/1800	0.161 290/1800	0.214 385/1800	0.064 116/1800
LLaVA	0.982 1768/1800	0.852 1534/1800	0.703 1265/1800	0.370 666/1800	0.147 264/1800	0.488 879/1800	0.014 25/1800
GPT-4V	0.969 1742/1797	0.888 1455/1799	0.889 1599/1800	0.680 1162/1710	0.238 428/1800	0.459 827/1800	0.265 473/1787
Gemini	0.993 1770/1782	0.838 1445/1724	0.922 1528/1658	0.754 1214/1611	0.271 473/1743	0.519 931/1794	0.343 600/1750

Table 2: Results of zero-shot performance across various domains on medical and molecule datasets.

Dataset	Camelyon17	HAM10000	NIH-Chest	COVID	DrugOOD_Assay	DrugOOD_Scaffold
Category	medical	medical	medical	medical	molecule	molecule
Prediction	tumor	skin diseases	lung disease	pneumonia types	bioassays	bioassays
Domain	hospital	hospital	hospital	hospital	assay	scaffold
#domains	5	4	2	2	81	12543
#classes	2	7	15	3	2	2
Examples						
CLIP	0.497	0.226	0.076	0.490	0.521	0.477
	894/1800	406/1800	137/1800	882/1800	924/1772	858/1800
LLaVA	0.508	0.160	0.089	0.420	0.521	0.477
	914/1800	288/1800	160/1800	756/1800	923/1772	859/1800
GPT-4V	0.513	0.341	0.084	0.313	0.488	0.514
	923/1799	548/1606	45/535	380/1216	414/848	647/1258
Gemini	0.532	0.335	0.119	0.515	0.490	0.508
	940/1766	572/1705	206/1729	926/1798	869/1772	914/1800

The comparative results, as detailed in Table 1, provide a quantitative analysis of GPT-4V's adaptability to distribution shifts across these domains. For natural image datasets characterized by diversity in artistic media, image repositories, everyday items, and various natural categories, GPT-4V generally outperforms CLIP and closely rivals or surpasses LLaVA. When faced with 1,800 random samples, its performance is very strong with 0.889 and 0.680 accuracies respectively. In the Appendix C, we also present the challenging scenario of failure cases where CLIP struggled, GPT-4V exhibits resilience, particularly in Office-Home, DomainNet, and Fmow dataset, where GPT-4V significantly exceeds CLIP's performance. These findings underscore the robustness of GPT-4V in complex natural visual tasks. Please refer to Appendix C for more evidence and case studies.

Quantitatively, GPT-4V's performance in natural images is consistently strong. However, in medical and scientific images, its performance suggests that the model may benefit from domain-specific training, as shown in Table 2. Please refer to Appendix C for more discussions and case studies.

It is very interesting to compare GPT-4V with traditional domain generalization methods to indicate the research direction of the domain generalization community. As shown in Table 5 in the Appendix C, GPT-4V's performance is notable in the DomainBed benchmark. It achieves superior zero-shot generalization capabilities, surpassing not only CLIP and LLaVA but also traditional approaches like ERM, MMD, and CORAL. With the highest average accuracy of 0.777 across various domains, GPT-4V establishes itself as a leading model, indicating its potential for versatile applications.

2.2 Adaptability to Controlled Data Perturbations

In assessing GPT-4V's adaptability to controlled data perturbations, specifically Gaussian noise and style changes introduced by ControlNet (Zhang et al., 2023), we observe that GPT-4V adeptly manages distribution shifts in various datasets, reflecting its strong zero-shot generalization abilities.

When examining domain shifts generated by ControlNet, as detailed in Table 3, GPT-4V continues to exhibit impressive generalization capabilities. For random samples in PACS_unseen and VLCS_unseen, the model achieves accuracies over 93%, and even in the more challenging Office-Home_unseen, it sustains a strong performance with 75.5% accuracy. GPT-4V's effectiveness is particularly highlighted in handling failure cases where CLIP exhibited no success; it achieves a remarkable accuracy rate of 61.1% in Office-Home_unseen.

The introduction of Gaussian noise to datasets such as PACS, VLCS, and Office-Home, which are believed not to be part of GPT-4V's pre-training, offers insights into the model's robustness to visual perturbations. As shown in Table 7 in the Appendix D, GPT-4V maintains high performance across all perturbed datasets, suggesting an inherent resilience to such noise. More discussions and case studies can be found in the Appendix. Overall, GPT-4V demonstrates a high degree of adaptability to controlled data perturbations, affirming its potential for robust applications in diverse environments.

Table 3: Main results of zero-shot performance across distribution shifts created by ControlNet.

Dataset	PACS_unseen	VLCS_unseen	Office-Home_unseen	PACS_unseen	VLCS_unseen	Office-Home_unseen
	random samples			***failure cases***		
CLIP	0.992	0.924	0.722	0.000	0.000	0.000
	1786/1800	1633/1768	1299/1800	0/16	0/135	0/180
LLaVA	0.996	0.962	0.618	0.813	0.726	0.250
	1793/1800	1700/1768	1113/1800	13/16	98/135	45/180
GPT-4V	0.989	0.932	0.755	0.875	0.880	0.611
	731/739	1096/1176	935/1238	14/16	117/133	110/180
Gemini	0.995	0.942	0.794	0.733	0.770	0.579
	1763/1772	1627/1728	1283/1615	11/15	97/126	95/164

2.3 Exploiting In-Context Learning for Domain Bridging

In the exploration of GPT-4V's in-context learning capabilities for domain bridging, the study assesses the model's adaptability to distribution shifts without traditional fine-tuning. Instead, it employs an innovative approach by conditioning GPT-4V with a few task demonstrations in its inputs, leveraging its ability to infer and apply patterns from limited data.

The study focuses on medical and scientific datasets, where GPT-4V previously underperformed, suggesting these domains were not well-represented in its pre-training data. The in-context learning setup involves providing GPT-4V with two examples from a source domain within these datasets and then evaluating its performance on a target domain. This method aims to simulate traditional domain adaptation/generalization in a more resource-efficient manner.

As shown in Figure 2, the implementation of in-context learning leads to consistent performance improvements across the selected datasets. For example, enhancements of 3.7% in Camelyon17 and a significant 16.67% in NIH-Chest are observed. These improvements underscore the effectiveness of in-context learning in enhancing the model's adaptability under distribution shifts. For more analyses and case studies please refer to the Appendix E.

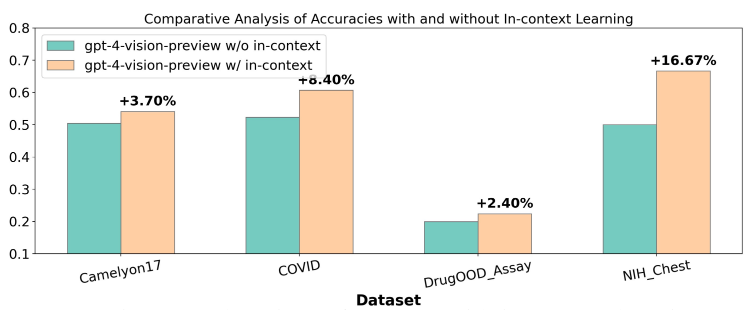


Figure 2: Improvements in target domain performance with in-context learning on GPT-4V across Camelyon17, COVID, DrugOOD_Assay and NIH_Chest datasets.

3 Observations and Discussion

Through rigorous evaluation and comparison with models across 13 diverse datasets, we have delineated the capability boundaries of GPT-4V, uncovering both its strengths and limitations as follows.

1. *Performance Across Domains:* GPT-4V shows robust performance in general but struggles in specialized domains like medicine and chemistry, indicating areas for improvement.
2. *Adaptability to Controlled Data Perturbations:* GPT-4V excels in handling novel and artificially perturbed data, showcasing superior generalization capabilities.
3. *In-context Learning Is an Effective Method:* Demonstrated effectiveness in enhancing adaptability to distribution shifts, suggesting a promising avenue for developing more sophisticated in-context learning methods to improve the robustness in specialized domains.
4. *Detail-Oriented Classification Rationale:* GPT-4V's nuanced understanding of image content surpasses LLaVA, emphasizing its advanced image classification capabilities.

5. *Higher Confidence in Predictions:* Exhibits higher confidence in its predictions, reflective of a confident decision-making process, with careful consideration in high-stakes contexts.
6. *Need for Domain-Specific Fine-Tuning:* Highlights the necessity for fine-tuning large foundation models in fields requiring specialized knowledge, where GPT-4V's accuracy could benefit from domain-specific data.
7. *Consistency in Challenging Samples:* Shows remarkable consistency and superior performance in challenging samples, particularly where other models like CLIP fail.
8. *Limitations in Applicability for Certain Tasks:* Struggles with tasks lacking semantic information of label, underscoring the need for task-specific adaptation or fine-tuning.

ACKNOWLEDGE

This work is supported by the National Natural Science Foundation of China (62176139), and the Major Basic Research Project of Natural Science Foundation of Shandong Province (ZR2021ZD15).

REFERENCES

Kartik Ahuja and David Lopez-Paz. A closer look at in-context learning under distribution shifts. *arXiv preprint arXiv:2305.16704*, 2023.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

Gilles Blanchard, Aniket Anand Deshmukh, Ürun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22(1): 46–100, 2021.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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.

Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34:21189–21201, 2021.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, pp. 440–457. Springer, 2022.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

DeepMind. Gemini - Google DeepMind — deepmind.google. https://deepmind.google/technologies/gemini/#hands-on, 2023. [Accessed 07-12-2023].

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

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

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020.

Sharut Gupta, Stefanie Jegelka, David Lopez-Paz, and Kartik Ahuja. Context is environment. *arXiv preprint arXiv:2309.09888*, 2023.

Zhongyi Han, Rundong He, Tianyang Li, Benzheng Wei, Jian Wang, and Yilong Yin. Semi-supervised screening of covid-19 from positive and unlabeled data with constraint non-negative risk estimator. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*, pp. 611–623. Springer, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021a.

Ruifeng Hu, Haodong Xu, Peilin Jia, and Zhongming Zhao. Kinasemd: kinase mutations and drug response database. *Nucleic Acids Research*, 49(D1):D552–D561, 2021b.

Zhuo Huang, Chang Liu, Yinpeng Dong, Hang Su, Shibao Zheng, and Tongliang Liu. Machine vision therapy: Multimodal large language models can enhance visual robustness via denoising in-context learning. *arXiv preprint arXiv:2312.02546*, 2023.

Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Lanqing Li, Long-Kai Huang, Tingyang Xu, Yu Rong, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8023–8031, 2023.

Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9619–9628, 2021.

Reto Knutti, Reinhard Furrer, Claudia Tebaldi, Jan Cermak, and Gerald A Meehl. Challenges in combining projections from multiple climate models. *Journal of Climate*, 23(10):2739–2758, 2010.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018b.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018c.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Akib Mashrur, Wei Luo, Nayyar A Zaidi, and Antonio Robles-Kelly. Machine learning for financial risk management: a survey. *IEEE Access*, 8:203203–203223, 2020.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8690–8699, 2021.

OpenAI. Gpt-4v(ision) system card. 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Chunjong Park, Anas Awadalla, Tadayoshi Kohno, and Shwetak Patel. Reliable and trustworthy machine learning for health using dataset shift detection. *Advances in Neural Information Processing Systems*, 34:3043–3056, 2021.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *arXiv preprint arXiv:2205.09739*, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Andrea Stocco and Paolo Tonella. Confidence-driven weighted retraining for predicting safety-critical failures in autonomous driving systems. *Journal of Software: Evolution and Process*, 34 (10):e2386, 2022.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Piotr Teterwak, Kuniaki Saito, Theodoros Tsiligkaridis, Kate Saenko, and Bryan A Plummer. Erm++: An improved baseline for domain generalization. *arXiv preprint arXiv:2304.01973*, 2023.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pp. 2097–2106, 2017.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

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. *arXiv preprint arXiv:2206.07682*, 2022a.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b.

Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6502–6509, 2020.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

A Our Approach in Exploring GPT-4V

Overall, our exploration of GPT-4V's adaptability to distribution shifts employs a nuanced approach, diverging from traditional fine-tuning due to the model's vast scale and the opaque nature of its pre-training data. We investigate through zero-shot generalization, assessing GPT-4V's natural adaptability across domains; response to data perturbations, examining robustness against artificially induced shifts; and in-context learning as a bridge across domains. This comprehensive evaluation strategy aims to illuminate GPT-4V's versatility in dynamic environments. To ensure a meaningful assessment, we've devised a sample selection strategy that balances diversity and informative value within OpenAI API's rate limits, including random sampling across domains and incorporating CLIP's failure instances for comparative analysis. Our approach extends to transforming classification tasks into a VQA format, designing prompts that facilitate a direct comparison of inference abilities between GPT-4V and other models, focusing on simplicity, uniformity, and insight into the models' reasoning processes.

A.1 How Do We Treat Distribution Shifts in This Work?

In the realm of machine learning, distribution shifts pose a formidable challenge, particularly when deploying models in real-world scenarios that differ from the training environment. Traditional approaches to this issue involve fine-tuning pre-trained foundation models on source domain data to adapt them to target domains. However, when it comes to massive models like GPT-4V, this conventional approach encounters significant hurdles. The vast scale of GPT-4V's architecture makes standard fine-tuning methods impractical, while the opacity nature of its pre-training data adds layers of uncertainty regarding its performance in novel scenarios. In response to these challenges, our study adopts a nuanced and multifaceted approach, aiming to thoroughly examine GPT-4V's adaptability to distribution shifts. This involves employing three distinct evaluation strategies: (1) **Zero-shot Generalization:** In Section C, we evaluate GPT-4V's inherent zero-shot generalization capabilities. Similar to models like CLIP, we assess GPT-4V's performance across different domains without prior tuning or exposure to relevant domain data, reflecting a purview into the model's natural adaptability. (2) **Response to Data Perturbations:** In Section D, our study investigates GPT-4V's robustness when faced with artificially induced shifts in data characteristics, focusing on domains where it shows initial high performance. (3) **In-Context Learning as a Domain Bridge:** In Section E, we assess GPT-4V's use of in-context learning to simulate conventional domain generalization paradigms, highlighting its ability to apply contextual understanding from the source domain to interpret data from the target one.

This multifaceted strategy is designed to illustrate GPT-4V's adaptability comprehensively, from its generalization capacity in comparison with baselines to its performance under artificially induced shifts and its ability to utilize contextual learning as a means to bridge domain gaps.

A.2 Sample Selection Guidance for GPT-4V Evaluation

To conduct a meaningful evaluation of GPT-4V within the constraints of the OpenAI API's rate limits, we have devised a sample selection strategy that prioritizes both diversity and informative value. Our selection process is guided by the following principles.

Random Sampling for Diversity. Our process begins with a random selection of samples from each class across all domains within our 13 datasets, intending to capture the inherent diversity of each domain, reflecting varying complexities and content types. To ensure comprehensive coverage, we employ two distinct sample sizes for each dataset: 180 and 1800. This decision aligns with OpenAI's revised rate limit policies, increasing daily requests from 100 to 500 as of December 2, 2023. Consequently, our sampling strategy, constrained by the limits of 100 and 500 requests per day for each account, strategically includes approximately 180 and 1800 random selections. Although these numbers might appear limited for certain datasets, they represent a balance between operational efficiency and practical feasibility under existing constraints. Notably, our preliminary findings indicate a consistent performance trend when comparing the two sample sizes, as shown in Tables 4 and 6. Our goal is to minimize selection bias and provide a thorough evaluation of GPT-4V's performance across a broad spectrum of data.

Prompt:

```
Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [dog, elephant, giraffe, guitar, horse, house, person]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.
```

Figure 3: An illustration of a structured prompt format used in the PACS dataset, showcasing a specific approach for image-based questioning and response formatting. The format includes a question about the image's content, a list of answer choices, and a template for answering, including an answer, confidence score, and the reasoning process.

Inclusion of Failure Cases From CLIP. To further enrich our evaluation, we have deliberately chosen to incorporate 180 instances for each dataset, where the CLIP model exhibits underperformance. This focused selection is driven by a specific objective: to assess how GPT-4V handles challenges that have proven difficult for a well-established model like CLIP. By analyzing GPT-4V's performance in these particular scenarios, we aim to gain deeper insights into its relative strengths and adaptability compared to CLIP. It is noteworthy that failure cases are sourced from CLIP due to its established role as a baseline model, particularly noted for its zero-shot robustness against distribution shifts. While a similar analytical approach using LLaVa's failure cases presents a valuable avenue for future research, it remains outside the scope of our current study.

Recognizing the continuous evolution of foundation models, such as Gemini (DeepMind, 2023), the cases we have selected are designed to function as a benchmark for evaluating and tracking the adaptability of state-of-the-art foundation models to distribution shifts. This benchmark not only serves our current study but also contributes to the broader research community.

A.3 Prompt Designs

In transforming conventional classification tasks into a visual question answering (VQA) format, our focus has been to devise a prompt template that is straightforward yet sufficiently informative. This approach seeks to exploit the expressive capabilities of language, a strength evidenced in previous models such as GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020). Crucially, our prompt design is tailored to a fair comparison of the inference abilities of GPT-4V and LLaVA. Specifically, we have developed a prompt that pairs an image with a clear, direct question, such as 'What is in this image?' followed by a set of answer choices. This design is intended to maintain simplicity, focusing primarily on the model's ability to interpret and accurately respond to visual content. Moreover, GPT-4V and LLaVA are prompted not just to select an answer option but also to provide a confidence score and a rationale for the decision, enhancing the depth of our analysis.

As exemplified in Figure 3, our structured prompt serves several key purposes:

- **Simplicity:** By employing a straightforward template that contextualizes the image with basic question and answer choices, we ensure minimal complexity in the prompt design.
- **Uniformity:** The approach ensures consistency and standardization in the model's responses, which is vital for comparative analysis across diverse test scenarios.
- **Insight into Reasoning:** The inclusion of confidence scoring and rationale requirements leverages GPT-4V's ability to output the decision-making process, thereby providing valuable insights into its reasoning and improving the interpretability of its outputs.

B MORE OBSERVATIONS

- *General Performance Across Domains:* In Section C, across various domains, GPT-4V showcased robust performance, particularly evidencing resilience to natural image distribution shifts. Nevertheless, its proficiency waned in more specialized fields like medicine and chemistry, signaling potential areas for enhancement. This was notably apparent in datasets such as Camelyon17, NIH-Chest, DrugOOD Assay, and DrugOOD Scaffold, where GPT-4V's classification outcomes resembled random guesses, as detailed in Table 6. This pattern suggests a need for targeted improvements in these domain-specific contexts.
- *Adaptability to Controlled Data Perturbations:* The experiments in Section D utilizing ControlNet-generated and random noise-induced data distributions presented GPT-4V with entirely novel domains, distinct from those involved in its pretraining phase. This setup rigorously tests the model's generalization capabilities in handling out-of-distribution scenarios. As demonstrated in Table 3, GPT-4V almost surpassed other methods in its performance, excelling particularly with challenging samples and in situations where CLIP encountered failures. These results underscore GPT-4V's exceptional stability and reliability when confronted with controlled perturbations and novel data distributions, highlighting its robust generalization abilities.
- *In-context Learning Is an Effective Method:* The experiments detailed in Section E illuminate the efficacy of in-context learning in enhancing GPT-4V's adaptability to distribution shifts. Notably, in the case studies depicted in Figure 5, GPT-4V demonstrates its capability to accurately identify the class of pathological images by discerning differences compared to two source images. This adaptability was consistently mirrored across four distinct datasets, reinforcing the utility of in-context learning strategies in navigating distribution shifts. Looking forward, there is a promising avenue for developing more sophisticated in-context learning methods, aiming to further bolster GPT-4V's robustness across diverse data distributions.
- *Detail-Oriented Classification Rationale:* The classification rationale provided by GPT-4V reflects a nuanced and detailed understanding of image elements, illustrating its sophisticated content comprehension. For instance, as exemplified in Figure 26, GPT-4V's capability outshines that of LLaVA by accurately recognizing distinct characteristics such as a robust body, short tail, and tufted ears. These instances clearly demonstrate GPT-4V's advanced ability to discern and articulate finer details in images, further reinforcing its superiority in complex image classification tasks under distribution shifts.
- *Higher Confidence in Predictions:* GPT-4V consistently displayed higher and more justified confidence levels in its predictions, indicative of a confident and precise decision-making process. As illustrated in Figure 26, GPT-4V's detail-oriented classification rationale contributes to its generating higher confidence scores compared to LLaVA. For instance, in Figure 17, GPT-4V achieves a peak confidence score with a descriptive analysis: "The image shows a metal kettle with a spout, handle, and thermometer on the top, which is a common design for a kettle used to heat water." Conversely, in medical imaging scenarios, such as depicted in Figure 30, GPT-4V's confidence scores are more moderate, often accompanied by recommendations for further clinical testing, reflecting a prudent approach in high-stakes contexts.
- *Need for Domain-Specific Fine-Tuning:* GPT-4V's performance in fields requiring specialized knowledge, such as medicine, chemistry, and biology, highlights the need for further fine-tuning using domain-specific data. While GPT-4V often provides rational and contextually appropriate reasoning, it can still yield incorrect classifications or diagnoses. A case in point is Figure 11, where GPT-4V accurately describes an image labeled as a guitar, stating that "the image displays a stylized depiction of a guitar ... leading to high confidence in this identification," yet it incorrectly classifies the image as a person. This example underscores the critical need for domain-specific fine-tuning, especially in areas where precision and reliability are paramount. Incorporating domain-specific knowledge and data into GPT-4V could substantially improve its accuracy, ensuring that its sophisticated reasoning consistently aligns with accurate contextual interpretations and decisions.
- *Consistency in Challenging Samples:* GPT-4V showcased remarkable consistency in handling challenging samples, particularly in scenarios where CLIP encountered errors. Its

performance was notably superior to that of LLaVA, exhibiting enhanced adaptability and precision. This is clearly evidenced in Tables 4 and 6, where, in instances of failure cases, GPT-4V almost outperforms both LLaVA and CLIP by a significant margin. These findings highlight GPT-4V's robustness and efficacy in dealing with complex samples, especially those involving significant distribution shifts.

- *Limitations in Applicability for Certain Tasks:* GPT-4V struggles with classification tasks when labels lack semantic information. This limitation becomes evident in scenarios such as activity identification tasks involving chemical molecular structures. In these cases, where sample labels are simply 'active' or 'inactive,' both GPT-4V and LLaVA tend to perform no better than random guessing. The provided reasoning, such as "The image shows a chemical structure, which does not have an active or inactive state in the context of physical motion or activity," as highlighted in Table 6 and Figure 32, reveals a gap in context comprehension. Similarly, tasks with numerical labels also pose a challenge for GPT-4V's zero-shot classification capabilities. These findings underscore the need for additional adaptation or fine-tuning for downstream tasks that involve non-semantic labels.

C Zero-shot Generalization Across Varied Domains

This section delineates our primary findings on the zero-shot generalization capabilities of GPT-4V in the context of distribution shifts, as enumerated in Table 4 and 6. We compare the performance of GPT-4V with baseline models such as CLIP[*] and LLaVA[†], highlighting its effectiveness and limitation across a variety of domains. Our investigation categorizes the datasets into three distinct groups: natural visuals, medical images, and molecular images. For each category, we first provide an overview of the collective results, showcasing GPT-4V's generalization performance. This is followed by in-depth case studies, where we delve into specific instances to uncover nuanced insights about the model's performance in diverse and challenging scenarios.

C.1 Natural Images

C.1.1 Task Introduction

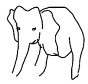
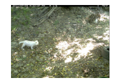
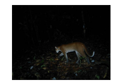
The category of natural visuals encompasses an extensive array of real-world imagery, capturing the myriad facets of nature and everyday life. This domain is characterized by its inherent diversity and complexity, presenting scenes and objects that are commonly encountered in daily experiences.

In our study, we examine the following natural datasets, each with its distinct characteristics and challenges:

- **PACS (Li et al., 2017):** Comprising images from four different styles - art painting, cartoon, photo, and sketch - this dataset challenges models to generalize across artistic mediums, testing their ability to recognize the same objects in vastly different visual representations.
- **VLCS (Fang et al., 2013):** This dataset is a collection from four different image repositories. It poses a challenge in terms of variations in image quality, lighting, and backgrounds, requiring robust feature extraction for successful classification.
- **Office-Home (Venkateswara et al., 2017):** Featuring everyday objects from office and home environments, this dataset includes images from diverse categories such as Art, Clipart, Product, and Real World, offering a testbed for models to generalize across everyday items.
- **DomainNet (Peng et al., 2019):** Encompassing a variety of artistic styles and objects, DomainNet is a large-scale dataset that tests a model's ability to generalize across different visual domains and a vast array of object classes.
- **Fmow (Christie et al., 2018):** This dataset focuses on land use and land cover classification, presenting a challenge with its time-series satellite imagery, which includes temporal and regional variations.

[*]https://huggingface.co/openai/clip-vit-base-patch16
[†]https://huggingface.co/liuhaotian/llava-v1.5-13b

Table 4: Summary of zero-shot generalization performance across various natural datasets, showcasing the comparative results of GPT-4V (gpt-4-vision-preview) with CLIP (clip-vit-base-patch16) and LLaVA (llava-v1.5-13b) models.

Dataset		PACS	VLCS	Office-Home	DomainNet	Fmow	TerraIncognita	iWildCam
Category		natural	natural	natural	natural	natural	natural	natural
Prediction		animal species	animal species	everyday items	objects, creatures	land use	animal species	animal species
Domain		artistic media	image repositories	visual categories	artistic styles	time, region	camera trap	location
#domains		4	4	4	6	6	4	206
#classes		7	5	65	345	62	10	323
Examples								
		random samples (180 cases)						
CLIP	acc	0.967	0.833	0.800	0.572	0.111	0.194	0.067
	ratio	174/180	150/180	144/180	103/180	20/180	35/180	12/180
LLaVA	acc	0.994	0.894	0.650	0.306	0.128	0.539	0.000
	ratio	179/180	161/180	117/180	55/180	23/180	97/180	0/180
GPT-4V	acc	0.978	0.797	0.936	0.833	0.220	0.500	0.309
	ratio	175/179	141/177	160/171	135/162	39/177	90/180	55/178
Gemini	acc	0.983	0.871	0.963	0.910	0.333	0.483	-
	ratio	173/176	148/170	155/161	142/156	56/168	87/180	-
		random samples (1800 cases)						
CLIP	acc	0.961	0.808	0.778	0.582	0.161	0.214	0.064
	ratio	1730/1800	1455/1800	1400/1800	1048/1800	290/1800	385/1800	116/1800
LLaVA	acc	0.982	0.852	0.703	0.370	0.147	0.488	0.014
	ratio	1768/1800	1534/1800	1265/1800	666/1800	264/1800	879/1800	25/1800
GPT-4V	acc	0.969	0.888	0.889	0.680	0.238	0.459	0.265
	ratio	1742/1797	1455/1799	1599/1800	1162/1710	428/1800	827/1800	473/1787
Gemini	acc	0.993	0.838	0.922	0.754	0.271	0.519	0.343
	ratio	1770/1782	1445/1724	1528/1658	1214/1611	473/1743	931/1794	600/1750
		failure cases						
CLIP	acc	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	ratio	0/173	0/180	0/180	0/180	0/180	0/180	0/180
LLaVA	acc	0.751	0.517	0.406	0.128	0.083	0.517	0.011
	ratio	130/173	93/180	73/180	23/180	15/180	93/180	2/180
GPT-4V	acc	0.732	0.651	0.774	0.523	0.192	0.411	0.285
	ratio	120/164	112/172	127/164	78/149	32/167	74/180	51/179
Gemini	acc	0.848	0.650	0.860	0.736	0.266	0.458	-
	ratio	140/165	104/160	141/164	106/144	45/169	82/179	-

- **TerraIncognita (Beery et al., 2018):** Composed of wildlife images captured by camera traps in various locations, it tests models' abilities to recognize animal species across different environmental conditions and camera settings.
- **iWildCam (Beery et al., 2021):** The iWildCam dataset offers a unique challenge in the realm of wildlife conservation and ecological studies. Comprised of images captured by camera traps set up in diverse wilderness locations, it is tailored to evaluate the ability of models to identify and classify a wide range of animal species.

These datasets not only cover a wide range of natural scenes and objects but also introduce various types of distribution shifts, making them ideal for evaluating the zero-shot generalization capabilities of GPT-4V, in comparison with CLIP and LLaVA. Each dataset presents its unique set of challenges, from artistic style variations in PACS to environmental differences in TerraIncognita, providing a comprehensive testbed for assessing model robustness in natural settings. Table 4 firstly provides an overview of each natural dataset, detailing key aspects such as the type of prediction, domain characteristics, the number of domains and classes, and illustrative examples. This table serves as a quick reference to understand the diversity and scope of challenges posed by these datasets in our evaluation.

C.1.2 Comparative Accuracies Across Datasets and Domains

Table 4 outlines the accuracies and correct-to-total case ratios for three models (CLIP, LLaVA, and GPT-4V) across six natural datasets, incorporating both random samples and failure cases identified in CLIP. This subsection is dedicated to examining GPT-4V's zero-shot generalization abilities within natural datasets.

Method	Office-Home	PACS	DomainNet	TerraIncognita	VLCS	Avg.
MMD (Li et al., 2018c)	0.663	0.847	0.234	0.422	0.775	0.588
Mixstyle (Zhou et al., 2021)	0.604	0.852	0.340	0.440	0.779	0.603
GroupDRO (Sagawa et al., 2019)	0.660	0.844	0.333	0.432	0.767	0.607
IRM (Arjovsky et al., 2019)	0.643	0.835	0.339	0.476	0.785	0.616
CDANN (Li et al., 2018b)	0.658	0.826	0.383	0.458	0.775	0.620
DANN (Ganin et al., 2016)	0.659	0.836	0.383	0.467	0.786	0.626
MTL (Blanchard et al., 2021)	0.664	0.846	0.406	0.456	0.772	0.629
Mixup (Xu et al., 2020)	0.681	0.846	0.392	0.479	0.774	0.634
MLDG (Li et al., 2018a)	0.668	0.849	0.412	0.477	0.772	0.636
ERM (Vapnik, 1999)	0.676	0.842	0.440	0.478	0.773	0.642
SagNet (Nam et al., 2021)	0.681	0.863	0.403	0.486	0.778	0.642
SelfReg (Kim et al., 2021)	0.679	0.856	0.428	0.470	0.778	0.642
CORAL (Sun & Saenko, 2016)	0.687	0.862	0.415	0.476	0.788	0.645
mDSDI (Bui et al., 2021)	0.692	0.862	0.428	0.481	0.790	0.651
ERM + MIRO (Cha et al., 2022)	0.705	0.854	0.443	0.504	0.790	0.659
ERM + SWAD (Cha et al., 2021)	0.706	0.881	0.465	0.500	0.791	0.669
CORAL + SWAD (Cha et al., 2021)	0.713	0.883	0.468	0.510	0.789	0.673
DIWA (Rame et al., 2022)	0.728	0.890	0.477	0.519	0.786	0.680
ERM + MIRO + SWAD (Cha et al., 2021)	0.724	0.884	0.470	**0.529**	0.796	0.681
ERM++ (Teterwak et al., 2023)	0.747	0.898	0.508	0.512	0.780	0.689
CLIP (Radford et al., 2021)	0.778	0.961	0.582	0.214	0.808	0.669
LLaVA (Liu et al., 2023b;a)	0.703	0.982	0.370	0.488	0.852	0.679
GPT-4V (OpenAI, 2023)	0.889	0.969	0.680	0.459	**0.888**	0.777
Gemini (Team et al., 2023)	**0.922**	**0.993**	**0.754**	0.519	0.838	**0.805**

Table 5: **Zero-shot Generalization Performance of GPT-4V on DomainBed:** In the DomainBed benchmark for domain generalization, GPT-4V demonstrates superior zero-shot generalization capabilities, surpassing traditional approaches and marking a significant advancement in the field. The results highlight GPT-4V's effectiveness across diverse domains, showcasing its potential for robust and versatile applications.

GPT-4V's Performance in Random Samples: Focusing first on datasets with a large variety of domains and classes, such as Office-Home and DomainNet, GPT-4V demonstrates a notable capacity for generalization. Its high accuracy rates in Office-Home (0.889) and DomainNet (0.680) suggest a robust understanding and adaptability to diverse natural visuals, including a broad range of everyday items and varied artistic styles. Additionally, in uncommon datasets like Fmow and TerraIncognita, GPT-4V significantly surpasses CLIP's performance (0.238 vs 0.161 in Fmow and 0.459 vs 0.214 in TerraIncognita). In the PACS and VLCS datasets, all three models perform well, with accuracies exceeding 0.8. This consistency suggests that these domains may have been included in the pre-training data of these three models.

GPT-4V in Handling CLIP's Failure Cases: To assess GPT-4V's capabilities in more challenging scenarios, we examine its performance on CLIP's failure cases. In datasets with a diverse range of classes, such as DomainNet and Office-Home, GPT-4V shows remarkable resilience. For instance, in Office-Home, GPT-4V achieves an accuracy of 0.774, surpassing LLaVA's 0.406. Similarly, in DomainNet, GPT-4V records 0.523 accuracy, significantly higher than LLaVA's 0.128. This trend is also evident in Fmow, where GPT-4V's performance (0.192) markedly exceeds LLaVA's (0.083). These results indicate GPT-4V's robustness in handling complex and challenging visuals, even in scenarios where CLIP struggled.

GPT-4V's Performance Across Individual Domains: While Table 4 provides an overall view of the accuracies for the three models across various datasets, a more granular look at their performance in specific domains is essential for a comprehensive understanding. To this end, we have detailed comparative domain accuracies for each model within the PACS, VLCS, Office-Home, DomainNet, Fmow, and TerraIncognita datasets. These comparisons are illustrated in Figures 7, 8. These figures illuminate the relative strengths and weaknesses of each model across different domains within the datasets and help to understand the extent of GPT-4V's generalization capabilities and how it compares to CLIP and LLaVA in diverse contexts.

Highlighting GPT-4V's Superiority in DomainBed: In the context of DomainBed (Gulrajani & Lopez-Paz, 2020), the popular benchmark for domain generalization, Table 5 provides a clear illustration of the strides made by GPT-4V. It achieves unparalleled zero-shot generalization performance, significantly outpacing traditional domain generalization methods. Its exceptional performance across the board is indicative of its sophisticated understanding and the ability to adapt to new, unseen domains. This achievement is not just a reflection of GPT-4V's powerful architecture but also an indicator of its potential to revolutionize how models tackle the challenge of domain generalization.

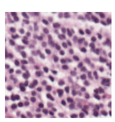
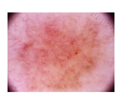
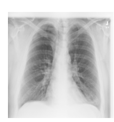
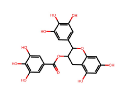
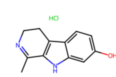
C.1.3 Case Demonstration

The diverse array of case studies presented in Figures 9, 10, 11, 13, 14, 15, 17, 18, 19, 21, 23, 24, 26, 27 and 28 showcase the adeptness of GPT-4V and LLaVA in navigating the challenges posed by different datasets, including PACS, VLCS, Office-Home, DomainNet, Fmow, and TerraIncognita. These examples not only demonstrate GPT-4V's proficiency in accurately recognizing natural distribution shifts in a zero-shot setting but also highlight its ability to adapt to various visual domains and object classifications. Additionally, Figures 12, 16, 20, 22, 25 and 29 provide insights into instances where GPT-4V does not perform optimally, shedding light on the model's limitations and areas for improvement.

A key observation emerging from these case studies is the nuanced capability of GPT-4V to discern intricate details within images. For instance, GPT-4V exhibits its adeptness at identifying textual elements in Figure 9. Figure 17 demonstrates a keen eye for specific features, such as the metallic nature and the spout of a kettle, highlighting its attention to detail. Furthermore, in Figure 26, GPT-4V distinguishes finer characteristics like a short tail and tufted ears in identifying a bobcat, a task that poses a challenge even for human observers.

C.2 Medical Images

Table 6: Main results of zero-shot generalization performance across distribution shifts on medical and molecule datasets. Specifically, CLIP refers to clip-vit-base-patch16, LLaVA refers to llava-v1.5-13b, GPT-4V refers to gpt-4-vision-preview.

Dataset	Camelyon17	HAM10000	NIH-Chest	COVID	DrugOOD_Assay	DrugOOD_Scaffold
Category	medical	medical	medical	medical	molecule	molecule
Prediction	tumor	skin diseases	lung disease	pneumonia types	bioassays	bioassays
Domain	hospital	hospital	hospital	hospital	assay	scaffold
#domains	5	4	2	2	81	12543
#classes	2	7	15	3	2	2
Examples						
	random samples (180 cases)					
CLIP	0.506	0.250	0.083	0.360	0.560	0.480
	91/180	45/180	15/180	36/100	56/100	96/200
LLaVA	0.508	0.160	0.089	0.450	0.560	0.480
	92/180	23/180	15/180	45/100	56/100	96/200
GPT-4V	0.518	0.291	0.072	0.354	0.445	0.500
	72/139	46/158	6/38	28/79	41/90	78/156
	random samples (1800 cases)					
CLIP	0.497	0.226	0.076	0.490	0.521	0.477
	894/1800	406/1800	137/1800	882/1800	924/1772	858/1800
LLaVA	0.508	0.160	0.089	0.420	0.521	0.477
	914/1800	288/1800	160/1800	756/1800	923/1772	859/1800
GPT-4V	0.513	0.341	0.084	0.313	0.488	0.514
	923/1799	548/1606	45/535	380/1216	414/848	647/1258
Gemini	0.532	0.335	0.119	0.515	0.490	0.508
	940/1766	572/1705	206/1729	926/1798	869/1772	914/1800
	failure cases					
CLIP	0.000	0.000	0.000	0.000	0.000	0.000
	0/180	0/180	0/180	0/100	0/100	0/132
LLaVA	0.028	0.083	0.083	0.510	0.010	0.000
	5/180	15/180	15/180	51/100	1/100	0/132
GPT-4V	1.000	0.308	0.102	0.543	1.000	1.000
	157/157	49/159	6/59	38/70	100/100	100/100

C.2.1 Task Introduction

We investigate the classification capabilities of different models in medical imaging applications under scenarios of distributional shifts. Distributional shifts are particularly common in the field of medical imaging, as changes in imaging technology, patient demographic characteristics, and disease manifestation can significantly alter the data distribution. Exploring the generalizability of the GPT-4 vision large model in medical image analysis tasks holds significant practical value.

In this part, we examine the following medical datasets, each with its distinct characteristics and challenges:

- **Camelyon17 (Bandi et al., 2018):** The dataset contains 450,000 patch samples, which were derived from 50 whole-slide images (WSIs) featuring breast cancer metastases in lymph node sections. These WSIs were sourced from five different hospitals in the Netherlands, contributing 10 WSIs each. Pathologists meticulously annotated each WSI to identify tumor regions, and these annotations were used to create segmentation masks. These masks, in turn, provided the basis for assigning labels to each individual patch in the dataset.
- **HAM10000 (Tschandl et al., 2018):** The dataset is a critical resource for research in skin lesion analysis, particularly focusing on generalization tasks. This dataset features a wide variety of dermatoscopic images, including numerous skin lesion types such as melanoma, basal cell carcinoma, and benign nevi. It is especially valuable for training and evaluating machine learning models on skin cancer detection and diagnosis. The diversity of images, sourced from different populations and equipment, makes HAM10000 ideal for studying and improving OOD generalization in medical imaging algorithms. This aspect is crucial for developing robust models capable of performing accurately across varied and unseen data, reflecting real-world clinical scenarios.
- **NIH-Chest (Wang et al., 2017):** The NIH Chest X-ray Dataset, a substantial medical imaging collection from the National Institutes of Health, is pivotal for research in out-of-distribution (OOD) generalization and distribution shift challenges in medical imaging. Comprising over 112,000 frontal-view X-ray images from more than 30,000 patients, this dataset is annotated with 14 common thoracic pathologies, such as pneumonia and lung nodules. Its vast and diverse array of patient images, captured under various clinical settings and conditions, provides an exceptional resource for developing and testing machine learning models, particularly in assessing and improving their robustness and performance in the face of distributional shifts and OOD data, which are common obstacles in real-world medical diagnostics.
- **COVID (Han et al., 2021):** This dataset serves as a resource for pneumonia detection, encompassing samples of normal cases, typical pneumonia, and COVID-19 pneumonia. The data, sourced from various hospitals due to collection methodologies, exhibit distributional shifts. We utilize this dataset to assess model performance in pneumonia detection tasks under conditions of distributional shift, reflecting real-world variations in medical data collection and patient demographics.

These datasets encompass a diverse array of medical scenarios and tasks, while also presenting a variety of distribution shifts. This diversity positions them as prime candidates for assessing the zero-shot generalization abilities of the GPT-4V model, with comparative analysis against CLIP and LLaVA. Table 6 offers a comprehensive overview of each dataset, highlighting crucial elements like prediction types, domain specifics, the range of domains and classes, along representative examples.

C.2.2 Comparative Accuracies Across Datasets and Domains

Table 6 outlines the accuracies and correct-to-total case ratios for three models (CLIP, LLaVA, and GPT-4V) across four medical datasets, incorporating both random samples and failure cases identified in CLIP. This subsection is dedicated to examining GPT-4V's zero-shot generalization abilities within medical datasets.

GPT-4V's Performance in Random Samples: According to Table 6, it is observed that the performance of GPT-4V, CLIP, and LLaVA on medical image classification tasks is quite average. For instance, on the Camelyon17 dataset, the performances of GPT-4V, CLIP, and LLaVA are 0.518, 0.506, and 0.511, respectively. This suggests that the data from these datasets may not have been present in the training sets of these three models, highlighting a potential gap in their pre-training data and indicating the need for further model training or adaptation to improve performance in these specific medical imaging tasks.

GPT-4V in Handling CLIP's Failure Cases: To assess GPT-4V's capabilities in more challenging scenarios, we examine its performance in CLIP's failure cases. On the HAM10000 dataset, GPT-4V achieved an accuracy of 0.308, surpassing LLaVa's 0.083. There were also varying degrees of accuracy improvements on the NIH-Chest and COVID datasets. These results demonstrate GPT-4V's robustness in handling complex and challenging visual tasks, maintaining stable performance even in scenarios where CLIP struggled.

C.2.3 Case Demonstration

The diverse array of case studies presented in Figures 30 and 31 showcase the adeptness of GPT-4V and LLaVA in navigating the challenges posed by different datasets, including HAM10000, NIH-Chest, and COVID.

C.3 Scientific Images

C.3.1 Task Introduction

Our research investigates the performance of various computational models in scientific fields, with a focus on predicting molecular properties amid distributional shifts due to variations in scaffolds and assays. Such shifts, resulting from changes in molecular scaffolds and assay conditions, profoundly affect the nature of scientific datasets. Assessing how advanced models like GPT-4 can adapt to these variations is vital for enhancing their predictive accuracy and reliability in the dynamic landscape of molecular science, where the intricate interplay of molecular structure and assay environments shapes data diversity and complexity.

In this part, we examine the following scientific datasets, each with its distinct characteristics and challenges:

DrugOOD (Ji et al., 2023) is a comprehensive dataset curator and benchmarking tool specifically designed for AI-aided drug discovery (AIDD). It focuses on the critical challenge of drug-target binding affinity prediction, involving both macromolecules (protein targets) and small molecules (drug compounds). Unlike traditional fixed datasets, DrugOOD offers automated data curation with customizable scripts, rich domain annotations, realistic noise annotations, and robust benchmarking of state-of-the-art OOD algorithms. It is particularly useful for testing graph-based out-of-distribution learning problems, crucial in molecular data modeled as irregular graphs. DrugOOD_Assay and DrugOOD_Scaffold can be obtained by splitting the domains with assays and scaffolds.

- **DrugOOD_Assay (Ji et al., 2023):** In the DrugOOD_Assay, domains are delineated based on the assay. This means that samples generated from the same assay are classified into the same domain, reflecting the unique environmental conditions of each assay. Due to these varying conditions, activity values measured across different assays exhibit a natural distribution shift. Consequently, the model is challenged to perform on data from bioassay environments it has not previously seen, testing its ability to generalize and maintain accuracy in the face of diverse and novel assay environments.
- **DrugOOD_Scaffold (Ji et al., 2023):** In the DrugOOD_Scaffold dataset, the domains are defined based on different molecular scaffolds. Molecules with the same molecular scaffold are grouped into the same domain, following the approach outlined by (Koh et al., 2021; Hu et al., 2021b). This structuring emphasizes the importance for models to have the capability to generalize effectively to unseen domains that are characterized by novel scaffolds, thereby enabling accurate predictions across a broad spectrum of molecular structures.

These datasets encompass a diverse array of scientific scenarios, while also presenting a variety of distribution shifts. This diversity positions them as prime candidates for assessing the zero-shot generalization abilities of the GPT-4V model, with comparative analysis against CLIP and LLaVA. Table 6 offers a comprehensive overview of each dataset, highlighting crucial elements like prediction types, domain specifics, the range of domains and classes, along representative examples.

Prompt:

As a chemistry expert specializing in molecular structures and chemical properties, your task is to examine the provided image of a chemical molecular structure. Using your profound knowledge in this field, assess the arrangement and types of atoms, bonding patterns, and overall molecular configuration. Your expertise will enable you to identify key indicators within the molecular structure that suggest whether it is 'active' or 'inactive' in terms of chemical reactivity. Please focus on aspects such as bond types, atom positioning, and molecular geometry, which are critical in determining the activity status of the molecule. Your detailed analysis should lead to a conclusion about the molecule's potential reactivity or stability based on the observable features in the structure.
Question: What is in this image? \\
Choices: [active, inactive]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

Figure 4: An illustration of a structured prompt format used in the PACS dataset, showcasing a specific approach for image-based questioning and response formatting. The format includes a question about the image's content, a list of answer choices, and a template for answering, including an answer, confidence score, and the reasoning process.

C.3.2 Performance Across Datasets and Domains

The results show that, in both the DrugOOD_Assay and DrugOOD_Assay datasets, GPT-4V, CLIP, and LLaVA failed. They were ineffective in accurately predicting the categories of molecules. The reasons for their failures could be attributed to three main factors: First, the complexity of the scientific task. Second, these datasets were not included in the training sets of these three models. Third, the ambiguity in data labeling, for instance, the labels 'inactive' and 'active' in scientific datasets are different from natural dataset labels like 'elephant' or 'bike'. The use of 'inactive' and 'active' as class labels is more ambiguous and lacks specific meaning. In conclusion, it is understandable that the zero-shot classification capabilities of these three models are poor.

C.3.3 Prompt Engineering Trick

This study explores the significant role of the Prompt Engineering Trick in enhancing performance in scientific image classification tasks. Specifically, we applied this technique in the task of chemical structure-activity classification, achieving a notable improvement in classification accuracy from 51.4% to 52.5%. This approach involves introducing meticulously designed prompts, such as instructing the model to analyze molecular structure images in the role of a chemistry expert, as shown in Figure 4. We required the model to not only identify atomic arrangements and bonding patterns in the images but also to interpret the overall configuration of the molecule to determine its chemical reactivity as either active or inactive. This method not only improved classification accuracy but also made the model's reasoning process more logical and interpretable. This research demonstrates that carefully designed prompts can significantly enhance the performance and understanding of machine learning models in specific tasks.

C.3.4 Case Demonstration

The representative case study presented in Figure 32 showcases the adeptness of GPT-4V and LLaVA in navigating the challenges. The results in Figures 32 show that GPT-4V does not perform well in predicting molecular properties. Although LLaVA can correctly predict the molecular properties, its reasoning is not convincing, suggesting that LLaVA's correct predictions are merely guesses without any solid basis. In contrast, although GPT-4V does not make accurate predictions, it does not provide a confidence level, and its reasoning is more logical. Therefore, to some extent, GPT-4V is more reliable than LLaVA.

D Adaptability to Controlled Data Perturbations

Based on the previous results in Figure 4, PACS, VLCS, and Office-Home are encountered in the GPT-4V's pre-training dataset. To investigate GPT-4V's capability in handling distribution shift, we adopt two approaches: (1) injecting Gaussian noise into PACS, VLCS, and Office-Home; (2) generating data with domain shift by ControlNet (Zhang et al., 2023). These newly obtained datasets naturally differ in distribution from the original datasets and are not seen during the pre-training process.

D.1 Gaussian Noise

Table 7: Main results of zero-shot generalization performance across distribution shifts created by adding Gaussian noise. Specifically, CLIP refers to clip-vit-base-patch16, LLaVA refers to llava-v1.5-13b, GPT-4V refers to gpt-4-vision-preview.

Dataset	PACS_gaussian	VLCS_gaussian	Office-Home_gaussian	PACS_gaussian	VLCS_gaussian	Office-Home_gaussian
		random samples			***failure cases***	
CLIP	0.961	0.799	0.741	0.000	0.000	0.000
	1729/1800	1439/1800	1334/1800	0/10	0/40	0/147
LLaVA	0.985	0.857	0.682	0.800	0.375	0.347
	1773/1800	1542/1800	1229/1800	8/10	15/40	51/147
GPT-4V	0.972	0.810	0.874	0.900	0.405	0.715
	1750/1800	1043/1287	1550/1773	9/10	15/37	93/130

D.1.1 Comparative Accuracies Across Datasets and Domains

Table 7 outlines the accuracies and correct-to-total case ratios for three models (CLIP, LLaVA, and GPT-4V) across PACS_gaussian, VLCS_gaussian, and Office-Home_gaussian, incorporating both random samples and failure cases identified in CLIP. This subsection is dedicated to examining GPT-4V's zero-shot generalization abilities within datasets with distribution shifts.

GPT-4V's Performance in Random Samples: Focusing initially on datasets encompassing a broad range of domains and categories, like Office-Home_gausssion, GPT-4V showcases remarkable generalization capabilities. Its impressive accuracy rate of 87.4% in Office-Home_gausssion is a testament to GPT-4V's adeptness in managing distribution shifts, especially those with Gaussian noise. In the PACS_Gaussian dataset, all three models exhibit strong performance, each surpassing an accuracy rate of 95%. This uniformity in performance hints that PACS_gausssion might have been a part of the foundational training data for these models.

GPT-4V in Handling CLIP's Failure Cases: To evaluate GPT-4V's performance in more challenging scenarios, we examined its response to cases where CLIP had failed. In datasets with a wide range of categories, such as Office-Home_gausssion, GPT-4V demonstrated significant resilience. For instance, in Office-Home_gausssion, GPT-4V achieved an accuracy rate of 71.5%, surpassing LLaVA's 34.7%. In both PACS_gausssion and VLCS_gausssion datasets, GPT-4V consistently outperformed LLaVA. These results highlight GPT-4V's robustness in handling complex and challenging visual scenarios, even in situations where CLIP encountered difficulties.

D.1.2 Case Demonstration

The diverse array of case studies presented in Figures 34, and 33 showcase the adeptness of GPT-4V and LLaVA in navigating the challenges posed by different datasets, including PACS_gaussian, Office-Home_gaussian, and VLCS_gaussian. These examples not only demonstrate GPT-4V's proficiency in accurately recognizing natural distribution shifts under Gaussian noise incorporation but also highlight its ability to adapt to various visual domains and object classifications.

D.2 Style Change with ControlNet

D.2.1 Comparative Accuracies Across Datasets and Domains

Table 3 outlines the accuracies and correct-to-total case ratios for three models (CLIP, LLaVA, Gemini, and GPT-4V) across PACS_unseen, VLCS_unseen, and Office-Home_unseen, incorporating both random samples and failure cases identified in CLIP. This subsection is dedicated to examining GPT-4V's zero-shot generalization abilities within datasets with domain shift created by ControlNet.

GPT-4V's Performance in Random Samples: Focusing initially on datasets encompassing a broad range of domains and categories, like Office-Home_unseen, GPT-4V showcases remarkable generalization capabilities. Its impressive accuracy rate of 75.5% in Office-Home_unseen is a testament to GPT-4V's adeptness in managing distribution shifts created by ControlNet. In the PACS_unseen and VLCS_unseen, all three models exhibit strong performance, each surpassing an accuracy rate of 90%. This uniformity in performance hints that PACS_unseen and VLCS_unseen might have been a part of the foundational training data for these models.

GPT-4V in Handling CLIP's Failure Cases: To evaluate GPT-4V's performance in more challenging scenarios, we examined its response to cases where CLIP had failed. In datasets with a wide range of categories, such as Office-Home_unseen, GPT-4V demonstrated significant resilience. For instance, in Office-Home_unseen, GPT-4V achieved an accuracy rate of 39.6%, surpassing LLaVA's 22.3%. In both PACS_unseen and VLCS_unseen datasets, GPT-4V consistently outperformed LLaVA. These results highlight GPT-4V's robustness in handling complex and challenging visual scenarios, even in situations where CLIP encountered difficulties.

D.2.2 Case Demonstration

The diverse array of case studies presented in Figure 36, 35, and 37 showcase the adeptness of GPT-4V and LLaVA in navigating the challenges posed by different datasets, including PACS_unseen, Office-Home_unseen, and VLCS_unseen. These examples not only demonstrate GPT-4V's proficiency in accurately recognizing natural distribution shifts created by ControlNet incorporation but also highlight its ability to adapt to various visual domains and object classifications. However, under certain complex samples, such as Figure 38, 39, and 40, both GPT-4V and LLaVA still have their limitations. They are prone to being misled by irrelevant factors in the image, leading to incorrect predictions.

E Exploiting In-Context Learning for Domain Bridging

Addressing distribution shifts traditionally involves fine-tuning pre-trained foundational models with source domain data to facilitate effective adaptation to target domains. While this approach can be effective, it often requires significant computational resources and time, especially for large foundational models (Hu et al., 2021a). Against this backdrop, our research shifts focus to the exploration of in-context learning capabilities of large multimodal models, with a specific emphasis on GPT-4V. This approach presents a novel method for simulating traditional domain generalization paradigms.

In-context learning, as defined by GPT-3 (Brown et al., 2020), involves conditioning the model on a set of natural language instructions alongside a few task demonstrations. The model is then expected to apply this learned context to complete further instances of the task, primarily through predicting subsequent sequences. This methodology leverages the model's inherent ability to infer and apply patterns from limited information without any parameter update, a significant difference from conventional fine-tuning techniques. This ability of large foundation models to demonstrate emergent capabilities through in-context learning has been increasingly recognized and highlighted in recent studies (Wei et al., 2022b; Ouyang et al., 2022; Wei et al., 2022a; Wang et al., 2022; Kojima et al., 2022). Our study aims to assess how effectively GPT-4V utilizes in-context learning to navigate distribution shifts across diverse domains (Ahuja & Lopez-Paz, 2023; Gupta et al., 2023).

E.1 In-context Setup

For our in-context learning exploration, we focus on the Camelyon17 (Bandi et al., 2018), COVID (Han et al., 2021), DrugOOD_Assay (Ji et al., 2023) and NIH-Chest (Wang et al., 2017) datasets. These datasets were chosen due to GPT-4V's previously observed underperformance, perhaps because the pre-training data distribution rarely includes scientific datasets like medical and protein. We wish the in-context learning that simulates conventional domain adaptation/generalization would enhance adaptability to certain tasks. In our experimental setup, we randomly select two classes within two domains of each dataset, designating them as source and target domains. From the source domain, we choose two representative examples for each class, like normal and typical pneumonia in the COVID dataset or normal and tumor in the Camelyon17 dataset, as illustrated in Figure 5. To demonstrate the potential of in-context learning as an effective approach for adapting large multimodal models to distribution shifts, we have intentionally limited our experiment to just one source domain and two examples. This decision is primarily driven by the constraints related to token cost. This setup emulates the concept of traditional out-of-distribution generalization but contrasts with it by leveraging the model's innate ability to adapt to new contextual information while maintaining its original parameterization (Brown et al., 2020).

Below, we illustrate an example of an in-context prompt applied to the Camelyon17 dataset. This dataset is distinguished by its binary classification system, encompassing two distinct classes: 'normal' and 'tumor'. In contrast to the basic prompt in Figure 3, we explicitly annotate the class labels for the two in-context examples provided to GPT-4V, i.e., 'The first image is normal and the second image is tumor'. Furthermore, the prompt's inquiry is subtly altered to 'What is the third image?', thereby aligning the model's focus with the specific task of classification based on the provided contextual examples. The response format template is set the same as the previous basic prompt.

Text Prompt with In-Context Examples:
Given the image, answer the following question using the specified format.
The first image is {class_1} and the second image is {class_2}.
Question: What is the third image?
Choices:['class_1', 'class_2'].
Please respond with the following format:
...

E.2 In-context Performance

In Figure 2, we illustrate the impact of in-context learning when applied to the baseline GPT-4V model, specifically within the target domain. This approach demonstrates consistent performance enhancements across four distinct datasets. In particular, the application of in-context learning yields improvements of 3.7%, 8.4%, 2.4%, and 16.67% for the Camelyon17, COVID, DrugOOD_Assay, and NIH_Chest datasets, respectively. These results highlight the potential of in-context learning in boosting model adaptability, especially in situations characterized by distribution shifts.

The observed variability in performance gains across these datasets suggests a correlation between the inherent task complexity and the unique data distributions of each dataset. This aspect of the results prompts further investigation into the nuances of in-context learning and its differential impact based on dataset characteristics.

In our experimental setup, two examples were randomly selected from the source domain for the in-context learning process. However, a more deliberate selection of in-context examples could potentially lead to even greater improvements in model performance (Huang et al., 2023). This possibility opens avenues for future research, where the strategic choice of in-context examples could be explored as a means to optimize the efficacy of in-context learning.

E.3 In-context Case Demonstration

This section showcases selected cases to demonstrate the enhancement of inference performance through in-context examples.

GPT-4V's Interpretation of In-context Examples: Figure 5 features a case study within the Camelyon17 dataset. The procedure includes presenting GPT-4V with two annotated images from a source domain (hospital_2): one denoted as 'normal' and the other as 'tumor'. These are followed by a test image from a different domain (hospital_3). Conditioned with this contextual information, GPT-4V effectively discerns between the regular, uniform tissue patterns in the ’normal’ image and the abnormal, irregular cell structures in the ’tumor’ image. It then applies this discernment to precisely classify the test image from hospital_3. This case exemplifies how GPT-4V employs in-context examples to bridge different domains, enhancing its interpretive accuracy.

The Impact of In-context Examples: Figure 6 explores the influence of in-context learning on GPT-4V’s performance in classifying chest X-ray images. The figure presents a comparative analysis of the model’s accuracy with and without in-context learning. Initially, GPT-4V incorrectly classifies a test image as ‘Pneumonia’ with a confidence score of 0.85, when no contextual information is provided. However, when conditioned with two in-context examples from the source domain, one labeled ’Pneumonia’ and the other ‘Normal,’ the model’s performance shifts markedly. With in-context learning, the model compares the third image with the first ’Pneumonia’ figure and accurately categorizes the same test image as ’Normal’ with an identical confidence score. This stark difference underscores the significant role that in-context learning plays in enhancing the model’s diagnostic precision, particularly in discerning subtle distinctions in medical imaging.

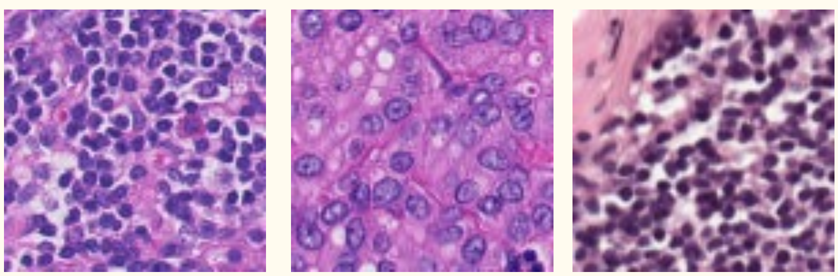
In-context Case

Prompt:

Given the image, answer the following question using the specified format. \\
The first image is normal and the second image is tumor.
Question: What is the third image? \\
Choices:['normal', 'tumor']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.



GPT-4V:

Answer Choice: normal
Confidence Score: 0.85
Reasoning:The **third image** presents tissue with cellular structures that appear to be regular and uniform, **similar to the first image labeled as normal**. There is no evident abnormal morphology typical of tumor tissue, such as irregular cell sizes, shapes, or disorganized arrangement, which is commonly observed **in the second image labeled as tumor**.

Figure 5: Demonstration of GPT-4V's inference process when exposed to in-context learning with examples from the Camelyon17 dataset. The experiment involves using two representative images from the source domain (hospital_2), one labeled 'normal' and the other 'tumor', followed by a test image from the target domain (hospital_3). GPT-4V, conditioned with these in-context examples, distinguishes between regular and uniform tissue patterns in the 'normal' image and abnormal, irregular cell sizes in the 'tumor' image. It then applies this contextual understanding to accurately infer the class of the test image from hospital_3. This process showcases GPT-4V's ability to leverage in-context cues for effective domain bridging.

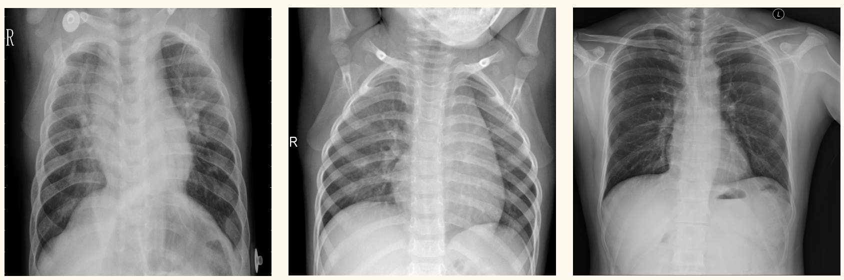
In-context Case

Prompt:

Given the image, answer the following question using the specified format. \\
The first image is Pneumonia and the second image is Normal.
Question: What is the third image? \\
Choices:['Pneumonia', 'Normal']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.



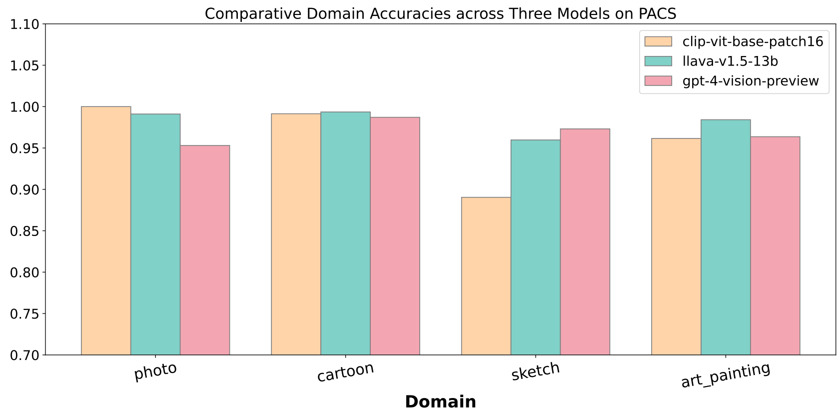
GPT-4V w/o in-context:

Answer Choice: Pneumonia
Confidence Score: 0.85
Reasoning: The X-ray image shows increased opacity in the lower lung fields which could indicate an infection or fluid, suggestive of pneumonia. The lung markings do not appear normal and there are areas that seem to be more opaque and consolidated, common indicators of pneumonia in a chest radiograph. However, a definitive diagnosis would require medical expertise and possibly additional information and testing.

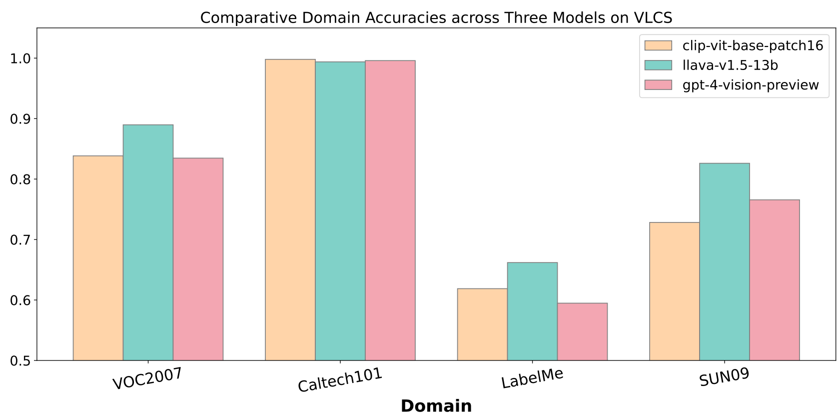
GPT-4V w/ in-context:

Answer Choice: Normal
Confidence Score: 0.85
Reasoning: The lung fields appear clear with no obvious consolidations or opacities that would suggest pneumonia. The heart size is normal, and there are no signs of pleural effusion. **Compared to the first image which was identified as Pneumonia**, this image lacks the characteristic signs such as lung consolidation and patchy opacities.

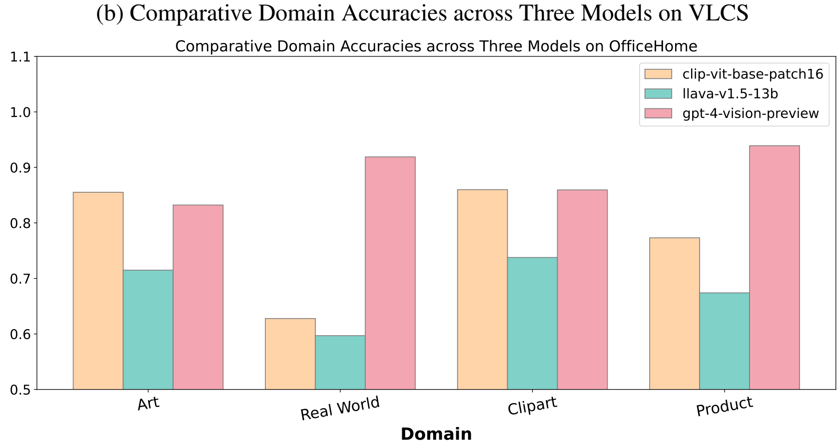
Figure 6: This figure illustrates GPT-4V's inference process under in-context learning, using selected examples from the COVID dataset. The process involves initially presenting two annotated images from the source domain: one labeled as 'Pneumonia' and the other as 'Normal.' Subsequently, GPT-4V is tasked with classifying a test image from the target domain. The model, having been conditioned with these specific in-context examples, evaluates and categorizes the test image by drawing comparisons to the previously presented pneumonia and normal X-ray images.



(a) Comparative Domain Accuracies across Three Models on PACS

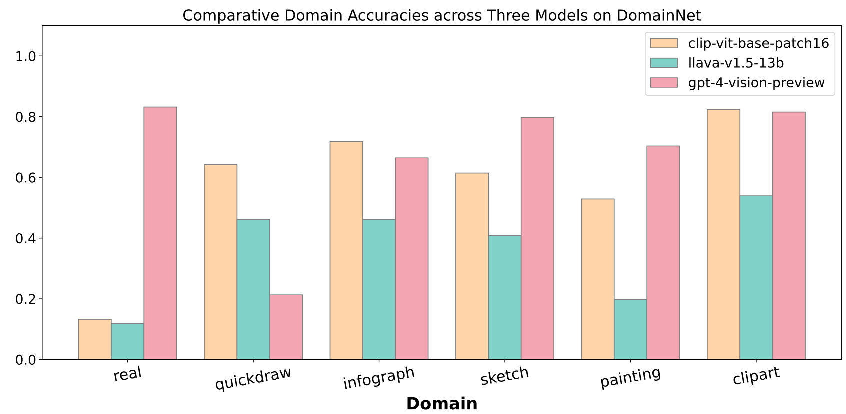


(b) Comparative Domain Accuracies across Three Models on VLCS

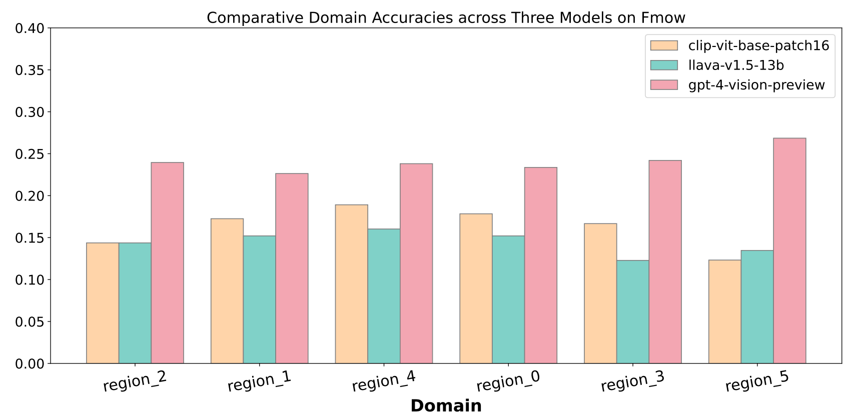


(c) Comparative Domain Accuracies across Three Models on OfficeHome

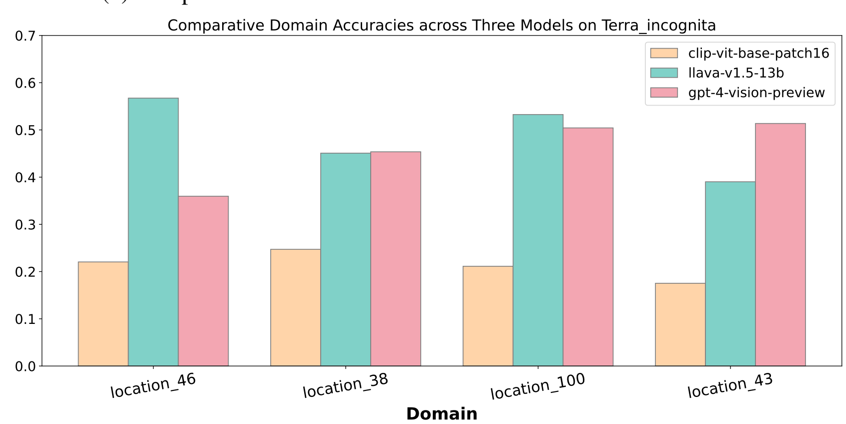
Figure 7: Comparative accuracies of three models across domains in the PACS, VLCS, Office-Home datasets.



(a) Comparative Domain Accuracies across Three Models on DomainNet



(b) Comparative Domain Accuracies across Three Models on Fmow



(c) Comparative Domain Accuracies across Three Models on TerraIncognita

Figure 8: Comparative accuracies of three models across domains in the DomainNet, Fmow, TerraIncognita datasets.

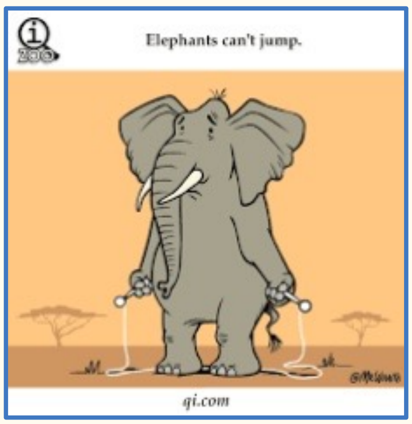
GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [dog, elephant, giraffe, guitar, horse, house, person]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.



GPT-4V:

Answer Choice: elephant
Confidence Score: 1
Reasoning: The image contains a **cartoon** illustration of an **elephant**, which is standing on a platform with springs attached to its feet. The caption **"Elephants can't jump"** also indicates that the focus of the image is an elephant.

LLaVA:

Answer Choice: elephant
Confidence Score: 0.9
Reasoning: The image features an **elephant with a trunk and tusks**, standing on a dirt field.

Figure 9: Natural Distribution Shift: Case 1 - Elephant category in the Cartoon Domain of PACS Dataset. In this instance, both GPT-4V and LLaVA are provided with the same text prompt alongside an image. Both models successfully generate an answer choice, a confidence score, and their reasoning. Notably, GPT-4V demonstrates a capability for detail recognition, accurately identifying the text 'Elephants can't jump.' in the image. This case exemplifies GPT-4V's advanced ability to discern and interpret finer details within visual inputs, compared to LLaVA.

GPT-4V: Incorrect | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [dog, elephant, giraffe, guitar, horse, house, person]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: giraffe
Confidence Score: 0.95
Reasoning: The image appears to be a **rough line** drawing of a giraffe. It is characterized by **a long neck**, **ossicones** (horn-like structures), and **facial features** that resemble a giraffe.

LLaVA:

Answer Choice: horse
Confidence Score: 0.8
Reasoning: The image is a drawing of a **horse**.

Figure 10: Natural Distribution Shift: Case 2 - Giraffe category in the Sketch Domain of PACS Dataset. This example showcases the performance of GPT-4V and LLaVA when given the same text prompt with an image depicting a giraffe in a sketch style. GPT-4V successfully identifies the giraffe, providing detailed reasoning and demonstrating a nuanced understanding of the image's content, such as long neck, horn-like structures. In contrast, LLaVA fails to correctly identify the giraffe, offering limited reasoning in its response.

GPT-4V: Incorrect | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [dog, elephant, giraffe, guitar, horse, house, person]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: person
Confidence Score: 0.95
Reasoning: The image displays a stylized depiction of a guitar, predominantly in the center, being played by a skeleton character. **The shape, strings, and tuning pegs are characteristic of a guitar**, leading to high confidence in this identification.

LLaVA:

Answer Choice: guitar
Confidence Score: None
Reasoning: The image shows a **skeleton playing a guitar**, which is the main focus of the image.

Figure 11: Natural Distribution Shift: Case 3 - Guitar category in the Art_painting Domain of PACS Dataset. While LLaVA accurately classifies the image, GPT-4V fails to identify the correct class. However, an interesting observation emerges in the rationale provided by GPT-4V. Despite the incorrect classification, GPT-4V articulates a highly reasoned and contextually relevant explanation, offering a detailed and accurate description of the ground_truth class label: guitar.

GPT-4V: Incorrect | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [dog, elephant, giraffe, guitar, horse, house, person]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

Ground-truth Answer: horse

GPT-4V:

Answer Choice: giraffe
Confidence Score: 1
Reasoning: The image contains illustrations of giraffes interspersed with trees and other small decorative elements, identifiable by their long necks and patterned bodies which are distinctive characteristics of giraffes.

LLaVA:

Answer Choice: dog
Confidence Score: None
Reasoning: The image is a pattern of dogs, and the dogs are in various poses and colors.

Figure 12: Natural Distribution Shift: Case 4 - Horse category in the Cartoon Domain of PACS Dataset. In this example, both GPT-4V and LLaVA incorrectly identify the subject in the image. The cartoon domain often features abstract styles where certain aspects of objects are exaggerated, as seen in the elongated necks of the horses in the image. GPT-4V incorrectly classifies the subject as a giraffe, likely influenced by the exaggerated neck feature. Compared to LLaVA, which provides limited reasoning, GPT-4V's rationale, though leading to an incorrect conclusion, is more detailed, noting the distinctive long neck as a key characteristic for its prediction.

GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['bird', 'car', 'chair', 'dog', 'person']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: chair
Confidence Score: 1
Reasoning: There are **multiple chairs** visible **in front of the cafe** in the image, **arranged around tables**. **No birds, cars, dogs, or persons are clearly identifiable** in the image.

LLaVA:

Answer Choice: chair
Confidence Score: 0.8
Reasoning: The image shows **a cafe with several chairs** outside.

Figure 13: Natural Distribution Shift: Case 5 - Chair category in the LabelMe Domain of VLCS Dataset. This case illustrates the proficiency of both GPT-4V and LLaVA models in accurately identifying multiple chairs within the scene. GPT-4V, in particular, stands out for its detailed and comprehensive description, offering nuanced insights that surpass the more straightforward analysis provided by LLaVA.

GPT-4V: Incorrect | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['bird', 'car', 'chair', 'dog', 'person']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: chair
Confidence Score: 1.0
Reasoning: There is a striped sofa in the image, which falls under the category of a chair as it is a piece of furniture designed for sitting. No birds, cars, dogs, or people are visible in the image.

LLaVA:

Answer Choice: None
Confidence Score: 0.9
Reasoning: The couch is the most prominent object in the image, taking up a significant portion of the living room.

Figure 14: Natural Distribution Shift: Case 6 - Chair category in the LabelMe Domain of VLCS Dataset. In this scenario, both GPT-4V and LLaVA models are presented with an image of a sofa/couch. GPT-4V demonstrates adaptability by categorizing the sofa as a type of chair, aligning with the limitations of the provided answer choices, and thus delivering an accurate classification. In contrast, LLaVA struggles to make the correct inference within the given constraints, highlighting a notable difference in their interpretative flexibility.

GPT-4V: Incorrect | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['bird', 'car', 'chair', 'dog', 'person']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: bird
Confidence Score: 0.95
Reasoning: There are multiple people visible walking on the sidewalk in the image.

LLaVA:

Answer Choice: person
Confidence Score: None
Reasoning: There are several people walking on the sidewalk and standing on the street corner, and some of them are carrying handbags.

Figure 15: Natural Distribution Shift: Case 7 - Person category in the LabelMe Domain of VLCS Dataset. In this instance, despite GPT-4V providing a logically sound reasoning process, it paradoxically arrives at an incorrect conclusion. This case highlights an intriguing aspect of GPT-4V's performance, where accurate analysis and reasoning do not always lead to the correct classification.

GPT-4V: Incorrect | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['bird', 'car', 'chair', 'dog', 'person']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

Ground-truth Answer: chair

GPT-4V:

Answer Choice: person
Confidence Score: 1
Reasoning: There is an individual seated at a table, **using a device which appears to be a phone or remote**, and this aligns with the 'person' answer choice.

LLaVA:

Answer Choice: person
Confidence Score: None
Reasoning: The image shows **a man sitting in a chair, eating food** and drinking wine. There are also wine glasses, a bottle, and a box of chocolates on the table.

Figure 16: Natural Distribution Shift: Case 8 - Chair category in the VOC2007 Domain of VLCS Dataset. This scenario illustrates the challenge faced by models like GPT-4V and LLaVA in accurately classifying images with multiple objects. Despite providing rational explanations, these models struggle to pinpoint the correct class when presented with complex scenes containing various elements.

GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['Bottle', 'Exit_Sign', 'Lamp_Shade', 'Postit_Notes', …, 'Speaker']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: Kettle
Confidence Score: 1
Reasoning: The image shows **a metal kettle with a spout, handle, and a thermometer on the top**, which is a common design for a kettle used to heat water.

LLaVA:

Answer Choice: Kettle
Confidence Score: 0.9
Reasoning: The image shows **a silver kettle with a black handle and a thermometer** on top.

Figure 17: Natural Distribution Shift: Case 9 - Kettle category in the Product Domain of Office-Home Dataset. In this case study, both GPT-4V and LLaVA models are tasked with responding to an identical text prompt accompanied by an image. It is noteworthy that GPT-4V demonstrates a more nuanced understanding, particularly in its ability to detail specific features such as the kettle's metallic nature and the presence of a spout. Additionally, GPT-4V enhances its answer with a summary that emphasizes typical design characteristics, thereby lending greater confidence to its response.

GPT-4V: Correct | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [‘Bottle’, ‘Exit_Sign’, ‘Lamp_Shade’, ‘Postit_Notes’, …, ‘Speaker’]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: Eraser
Confidence Score: 1
Reasoning: The image displays a hand holding a large pink eraser with the words "BIG MISTAKE" on it, and there are remnants of erased material, likely from a pencil, below it on the surface.\n

LLaVA:

Answer Choice: Marker
Confidence Score: 0.8
Reasoning: The image shows a person holding a pink object, which is a marker, and using it to write on a piece of paper.

Figure 18: Natural Distribution Shift: Case 10 - Analyzing the ’Eraser’ Category in the Art Domain of the Office-Home Dataset. This figure presents an intriguing instance where the depicted ’Eraser’ might be initially mistaken for a ’Marker’, a common perceptual challenge. GPT-4V remarkably identifies the correct object, utilizing cues from the text in the image, as well as the object’s size and color. Notably, GPT-4V correctly interprets the action of erasing, in contrast to LLaVA, which interprets the action as writing. This demonstrates GPT-4V’s advanced reasoning capabilities in distinguishing subtle contextual differences.

GPT-4V: Incorrect | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['Bottle', 'Exit_Sign', 'Lamp_Shade', 'Postit_Notes', …, 'Speaker']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: Table
Confidence Score: 1
Reasoning: The image clearly shows a **wooden chair** in the foreground, which is among the choices listed. There is no ambiguity in the image regarding the object\'s identity as a chair.

LLaVA:

Answer Choice: Chair
Confidence Score: 0.9
Reasoning: The image shows a wooden chair with a leather seat and back.

Figure 19: Natural Distribution Shift: Case 11 - Chair category in the Real World Domain of Office-Home Dataset. In this example, GPT-4V exhibits details and accuracy in its description of the image. Despite this, the model ultimately arrives at an incorrect classification.

GPT-4V: Incorrect | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['Bottle', 'Exit_Sign', 'Lamp_Shade', 'Postit_Notes', …, 'Speaker']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

Ground-truth Answer: Couch

GPT-4V:

Answer Choice: Chair
Confidence Score: 1
Reasoning: The image displays a large chair with red upholstery and a unique design featuring skeletal and demonic features. It matches the description of a chair by providing seating for individuals, although the style is highly thematic and ornate.

LLaVA:

Answer Choice: Chair
Confidence Score: 0.9
Reasoning: The image shows a red chair with a black metal frame and a red velvet seat.

Figure 20: Natural Distribution Shift: Case 12 - Couch category in the Art Domain of Office-Home Dataset. In this instance, both GPT-4V and LLaVA demonstrate detailed and accurate descriptions of the image, yet both models misclassify the object. This misclassification arises from the overlapping categories of 'couch' and 'chair' in the dataset, showcasing the challenge models face when distinct class labels share similarities. This case highlights the complexity models encounter in accurately categorizing objects within overlapping or closely related classes.

GPT-4V: Correct | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices:['teddy-bear', 'strawberry', 'spoon', 'skull', 'school_bus', 'rain', 'pizza', 'parrot', 'ocean', 'line', 'monkey', …, 'bridge', 'camera']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: finger
Confidence Score: 1
Reasoning: The image clearly shows an illustration of a human hand with **five fingers extended**, which corresponds with the 'finger**' choice from the provided list**.

LLaVA:

Answer Choice: hand
Confidence Score: 0.9
Reasoning: The image is a **white hand**.

Figure 21: Natural Distribution Shift: Case 13 - Analyzing the ’Finger’ Category in the Clipart Domain of the DomainNet Dataset. This figure showcases a challenging example where differentiation between ’Hand’ and ’Finger’ is crucial. GPT-4V accurately identifies the object as ’Finger’, recognizing the illustration of a human hand with all five fingers extended, which aligns with the ’Finger’ choice. In contrast, LLaVA identifies the object as ’Hand’, providing a less precise classification and a somewhat unusual description of a ’white hand’.

GPT-4V: Incorrect | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [‘Bottle’, ‘Exit_Sign’, ‘Lamp_Shade’, ‘Postit_Notes’, …, ‘Speaker’]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

Ground-truth Answer: cup

GPT-4V:

Answer Choice: coffee_cup
Confidence Score: 1
Reasoning: The image shows a mug typically used for coffee, with the phrase **"oh what a Beautiful morning"** and decorative floral patterns on its surface. This is consistent with the provided choice of "coffee_cup."

LLaVA:

Answer Choice: mug
Confidence Score: 0.9
Reasoning: The image features a white coffee mug with a floral design and the words **"Oh What a Beautiful Morning"** written on it.

Figure 22: Natural Distribution Shift: Case 14 - Analyzing the ’Cup’ Category in the Painting Domain of the DomainNet Dataset. Despite both GPT-4V and LLaVA present accurate OCR capability, neither GPT-4V nor LLaVA successfully distinguishes the correct category among these closely related concepts, leading to incorrect classifications. This scenario underscores the complexity inherent in nuanced visual recognition tasks, particularly when dealing with objects that share similar characteristics and uses.

GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['airport', 'airport_hangar', 'airport_terminal', 'amusement_park', 'aquaculture', ..., 'wind_farm', 'zoo']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: airport
Confidence Score: 0.9
Reasoning: The image shows a long, clear area that appears to be a dirt or **unpaved runway**, which is characteristic of **a small airport or airfield**. Structures that could be associated with **aircraft hangars or terminals** are adjacent to this runway. The layout is consistent with that of **a small, possibly regional, airport**.

LLaVA:

Answer Choice: airport
Confidence Score: 0.8
Reasoning: The image shows a large area of land with a road and buildings, and there is a clearing in the middle of the area. This clearing is **likely an airport**, as it is surrounded by buildings and infrastructure. The presence of a road and buildings also supports the idea that it is an airport, as these are **common features of airport locations**.

Figure 23: Natural Distribution Shift: Case 15 - Evaluating the 'Airport' Category in the Region_2 Domain of the Fmow Dataset. In this case study, both GPT-4V and LLaVA models are tasked with responding to an identical text prompt accompanied by an image. Each model correctly identifies 'airport' as the subject, with GPT-4V exhibiting a higher confidence score of 0.9, as opposed to LLaVA's 0.8. GPT-4V stands out for its detailed analysis, identifying key elements like runways, aircraft, and terminals, indicative of an airport. Remarkably, GPT-4V further distinguishes the airport as small or regional, showcasing its advanced reasoning and contextual interpretation abilities.

GPT-4V: Correct | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['airport', 'airport_hangar', 'airport_terminal', 'amusement_park', 'aquaculture', …, 'wind_farm', 'zoo']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: port
Confidence Score: 0.9
Reasoning: The image shows **large docking areas** with ships, as well as infrastructure such as **cranes and storage areas typically associated with a port**. The proximity to water with various piers extending into it is characteristic of ports, where cargo ships dock to load and unload goods.

LLaVA:

Answer Choice: airport
Confidence Score: 0.9
Reasoning: The image shows a large area with many buildings, including a runway, which is a characteristic of an airport.

Figure 24: Natural Distribution Shift: Case 16 - Evaluating the 'Port' Category in the Region_0 Domain of the Fmow Dataset. In this instance, GPT-4V accurately identifies the location as a port, citing the presence of docking areas, ships, cranes as key indicators. Its reasoning is thorough, focusing on specific port-related characteristics. Conversely, LLaVA incorrectly classifies the same image as an airport, referencing runways and buildings, but lacks the detailed analysis of maritime infrastructure present in GPT-4V's description.

GPT-4V: Incorrect | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['airport', 'airport_hangar', 'airport_terminal', 'amusement_park', 'aquaculture', …, 'wind_farm', 'zoo']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: water_treatment_facility
Confidence Score: 0.85
Reasoning: The structures in the image **resemble large tanks or basins** that are often characteristic of water treatment facilities, where water is processed in large containment areas for purification or sewage treatment purposes. There are clear geometric structures that are typical for such facilities.

LLaVA:

Answer Choice: dam
Confidence Score: 0.9
Reasoning: The image **shows a large body of water with a dam**, which is a structure built across a waterway to control the flow of water.

Figure 25: Natural Distribution Shift: Case 17 - Evaluating the 'Dam' Category in the Region_3 Domain of the Fmow Dataset. In this image, GPT-4V incorrectly identifies the scene as a water treatment facility, citing the presence of large tanks or basins typically found in such settings. Conversely, LLaVA correctly classifies the image as a dam, accurately recognizing the large body of water and the structure controlling its flow, with a confidence score of 0.9.

GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['bird', 'bobcat', 'cat', 'coyote', 'dog', 'empty', 'opossum', 'rabbit', 'raccoon', 'squirrel']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: bobcat
Confidence Score: 0.95
Reasoning: The animal in the image has physical features consistent with a bobcat, such as **a robust body, short tail, and tufted ears**. The **size, pattern, and color** of the coat match typical bobcat characteristics.

LLaVA:

Answer Choice: bobcat
Confidence Score: 0.8
Reasoning: The image shows **a small animal** walking on a dirt road, and it is described as a bobcat.

Figure 26: Natural Distribution Shift: Case 18 - Analyzing the 'Bobcat' Category in Location_46 Domain of the TerraIncognita Dataset. In this evaluation, GPT-4V and LLaVA models respond to a uniform text prompt accompanied by a wildlife image. Both models accurately identify a 'bobcat' as the subject. GPT-4V shows a higher confidence score of 0.95, compared to 0.8 by LLaVA. GPT-4V's reasoning is notable for its detailed analysis, focusing on distinctive physical features of the bobcat, such as a robust body, short tail, and tufted ears, which are challenging to discern even for humans. It also augments its response by detailing the size, pattern, and color of the bobcat, which are crucial for accurate identification. In contrast, LLaVA's identification is based on the general observation of a small animal, a criterion that could apply to multiple species, thereby lacking the specificity demonstrated by GPT-4V.

GPT-4V: Correct | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['bird', 'bobcat', 'cat', 'coyote', 'dog', 'empty', 'opossum', 'rabbit', 'raccoon', 'squirrel']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: coyote
Confidence Score: 0.75
Reasoning: The animal in the image has **a slender, elongated snout and an upright, bushy tail**, which are characteristic features of a coyote. The general body shape and size in the image are also consistent with that of a coyote. However, the image is monochromatic and grainy, which might obscure finer details needed for a more definitive identification.

LLaVA:

Answer Choice: cat
Confidence Score: 0.8
Reasoning: The animal in the image is a cat, as it is walking in the dark and appears to be **a small house cat**.

Figure 27: Natural Distribution Shift: Case 19 - Analyzing the 'Coyote' Category in Location_46 Domain of the TerraIncognita Dataset. In this image, GPT-4V accurately identifies the animal as a coyote, noting its slender, elongated snout and upright, bushy tail, and assigning a confidence score of 0.75. It carefully considers the monochromatic and grainy nature of the image that may obscure finer details. In contrast, LLaVA incorrectly classifies the animal as a cat with a confidence score of 0.8, failing to recognize the distinct features of a coyote.

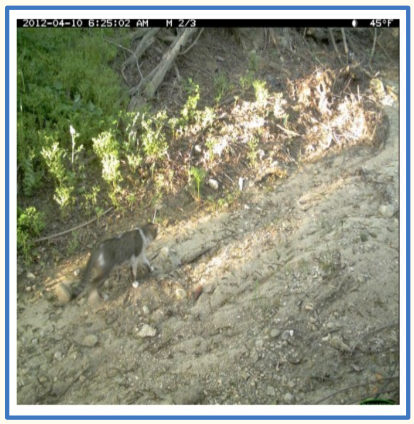
GPT-4V: Incorrect | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['bird', 'bobcat', 'cat', 'coyote', 'dog', 'empty', 'opossum', 'rabbit', 'raccoon', 'squirrel']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.



GPT-4V:

Answer Choice: bobcat
Confidence Score: 0.80
Reasoning: The animal in the image has a physical build and coat pattern that resemble those of a bobcat, with apparent tufted ears and **a short tail**. The environment and the solitary nature of the subject also align with the typical habitat and behavior of a bobcat.

LLaVA:

Answer Choice: cat
Confidence Score: 0.9
Reasoning: The image shows a cat walking on a dirt road.

Figure 28: Natural Distribution Shift: Case 20 - Analyzing the 'Cat' Category in Location_38 Domain of the TerraIncognita Dataset. In this case, LLaVA correctly identifies the animal as a 'cat' with a higher confidence score of 0.9, whereas GPT-4V, with a confidence score of 0.8, mistakenly identifies the animal as a 'bobcat'. The detailed reasoning of GPT-4V highlights its misclassification: it points to features such as a perceived short tail and tufted ears, typically characteristic of a bobcat, leading to its incorrect conclusion. This case illustrates the nuances and challenges of wildlife species recognition, particularly in distinguishing between visually similar animals in natural environments. The confidence score outputted by GPT-4V, despite its misclassification in this instance, can serve as a valuable metric, offering insights into the model's decision-making process and potentially guiding reliance on its conclusions.

GPT-4V: Incorrect | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['bird', 'bobcat', 'cat', 'coyote', 'dog', 'empty', 'opossum', 'rabbit', 'raccoon‘, 'squirrel']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

Ground-truth Answer: rabbit

GPT-4V:

Answer Choice: bobcat
Confidence Score: 0.95
Reasoning: The animal has the distinctive physical features of a bobcat, including **tufted ears, a short tail, and a stocky build**. The size and physical characteristics are consistent with those of a bobcat rather than the other listed animals.

LLaVA:

Answer Choice: bobcat
Confidence Score: None
Reasoning: The image shows **two animals** walking together, and the presence of a tree in the background suggests a natural environment. The animals are walking in the dirt, which is typical for bobcats. The fact that there are two animals walking together also supports the idea that they are bobcats, as they are known to be social animals.

Figure 29: Natural Distribution Shift: Case 21 - Analyzing the 'Rabbit' Category in Location_43 Domain of the TerraIncognita Dataset. This image presents a complex wildlife scenario where a bobcat is seen capturing and biting a rabbit. Both GPT-4V and LLaVA misinterpret the scene by identifying only the bobcat. This case underlines the intricacies of wildlife recognition, particularly when multiple animals interact in a single frame. The primary focus on the bobcat, while ignoring the rabbit, points to the nuanced challenges in accurately interpreting dynamic natural scenes.

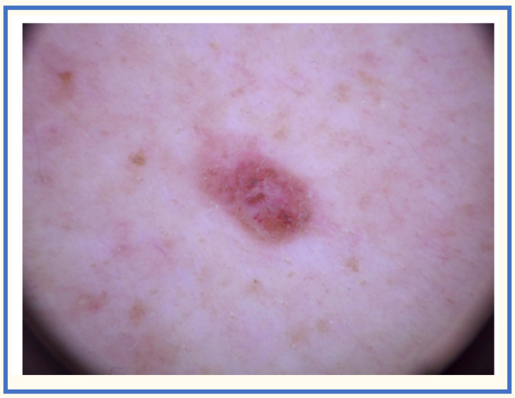
GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [‘actinic keratoses and intraepithelial carcinoma’, …, ‘melanoma’]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.



GPT-4V:

Answer Choice: basal cell carcinoma
Confidence Score: 0.7
Reasoning: The lesion in the image appears **pearly and translucent with what may be telangiectatic vessels, which are common characteristics of basal cell carcinoma**. However, without further clinical context or a diagnostic biopsy, it's not possible to be certain.

LLaVA:

Answer Choice: basal cell carcinoma
Confidence Score: 0.75
Reasoning: The image shows **a skin lesion that is red and brown in color, which is consistent with the appearance of basal cell carcinoma**. Additionally, the lesion is described as being “ugly” and “ugly looking,” which further supports the possibility that it is a skin cancer.

Figure 30: Medical Distribution Shift: Case 1 - Analyzing the ‘basal cell carcinoma’ Category in the vidir_modern Domain of the HAM10000 Dataset. In this case study, both GPT-4V and LLaVA models are tasked with responding to an identical text prompt accompanied by an image. Each model correctly identifies ‘basal cell carcinoma’ as the subject, with LLaVA exhibiting a higher confidence score of 0.75, as opposed to GPT-4V’s 0.7. GPT-4V stands out for its detailed analysis, identifying key elements like pearly, translucent, and telangiectatic vessels, indicative of basal cell carcinoma. LLaVA gives an analysis mainly in terms of color appearance.

GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['COVID19', 'Normal', 'Pneumonia']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: Normal
Confidence Score: 0.7
Reasoning: The image appears to be a standard chest X-ray **without obvious signs of pneumonia, such as localized opacities**, or signs typically associated with COVID-19, **such as bilateral peripheral ground-glass opacities**. However, without clinical context or additional diagnostic information, this interpretation may not be definitive.

LLaVA:

Answer Choice: Normal
Confidence Score: 0.95
Reasoning: The image shows a **baby's ribcage**, which is a normal part of a human body.

Figure 31: Medical Distribution Shift: Case 2 - Analyzing the 'Normal' Category in the hostial1 Domain of the COVID Dataset. In this case study, both GPT-4V and LLaVA models are tasked with responding to an identical text prompt accompanied by an image. Each model correctly identifies 'Normal' as the subject, with LLaVA exhibiting a higher confidence score of 0.95, as opposed to GPT-4V's 0.7. GPT-4V stands out for its detailed analysis, identifying key elements like localized opacities, and bilateral peripheral ground-glass opacities, indicative of a normal sample. LLaVA states that the image is of a baby's rib cage, but does not give a proper reason for why it is categorized as normal, despite the high confidence score of 0.95.

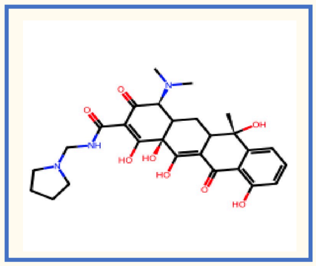
GPT-4V: Inorrect | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [‘active’, ‘inactive’]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.



GPT-4V:

Answer Choice: Inactive
Confidence Score: null
Reasoning: The image shows a **chemical structure**, which **cannot be categorized as 'active' or 'inactive' without additional context such as biological activity**, enzyme interaction, or the presence in an active/inactive state within a metabolic pathway. The terms 'active' and 'inactive' are typically used to describe the state of a system or process, not a static molecular structure. Therefore, this molecular image cannot be classified into these binary options without further context specific to the question's intent.

LLaVA:

Answer Choice: Active
Confidence Score: 0.9
Reasoning: The image shows a chemical structure with **a red and blue bond**, which indicates that it is an active compound.

Figure 32: Scientific Distribution Shift: Case 1 - Analyzing the ‘active’ Category in the ID-75 Domain of the DrugOOD_assay Dataset. In this case study, both GPT-4V and LLaVA models are tasked with responding to an identical text prompt accompanied by an image. GPT-4V incorrectly predicts without a confidence score while LLaVA correctly predicts with high confidence. Although the GPT-4V predicts error categories, it does not give high confidence. According to reason, GPT-4V can recognize this image as a chemical structure. The statement that it cannot be categorized as ‘active’ or ‘inactive’ without other contexts such as biological activity, enzyme interactions, or active/inactive states in metabolic pathways makes sense. The limitations of the DrugOOD dataset itself are also noted here, i.e., it is not reasonable to simply categorize the data as ‘active’ or ‘inactive’. Conversely, LLaVA, despite giving correct predictions, is very unreliable in its reasoning. It is wrong to classify them as ‘active’ or ‘inactive’ by the color of the bond.

GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['dog', 'elephant', ..., 'person']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: house
Confidence Score: 0.75
Reasoning: The image appears to depict **buildings**, one of which has **a spire or steeple, resembling structural architecture typically associated with a house or a church**. The image is low-resolution and abstract, making it difficult to ascertain complete details, **but there are clearly no animals, people, or guitars present**.

LLaVA:

Answer Choice: house
Confidence Score: None
Reasoning: The image shows a **house with a clock tower and a boat in the water**. The presence of the clock tower and the boat suggest that this is a house with a unique architectural feature, making it stand out from other houses. The house is also **located near a body of water**, which is a common feature in many houses.

Figure 33: Induced Distribution Shift: Case 1 - Analyzing the 'Fork' Category in the Product Domain of the Office-Home_gaussion Dataset. In this case study, both GPT-4V and LLaVA models are tasked with responding to an identical text prompt accompanied by an image. Both GPT-4V and LLaVA predicted correctly, but GPT-4V gave higher confidence levels as well as more detailed explanations such as tines, a handle, and several pointed prongs. These are the basic characteristics of a fork. However, the reason given by LLaVA is rough.

GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['Alarm_Clock', 'Backpack', …, 'Webcam']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: Fork
Confidence Score: 1.00
Reasoning: The image clearly displays a utensil with **tines** which is used for eating, commonly known as a fork. It matches the typical appearance of a fork with **a handle and several pointed prongs**.

LLaVA:

Answer Choice: Fork
Confidence Score: 0.90
Reasoning: The image shows a **fork** on a blue tablecloth.

Figure 34: Induced Distribution Shift: Case 2 - Analyzing the 'house' Category in the art_painting Domain of the PACS_gaussion Dataset. In this case study, both GPT-4V and LLaVA models are tasked with responding to an identical text prompt accompanied by an image. Both GPT-4V and LLaVA predicted correctly, however, GPT-4V gave a confidence level of 0.75 and LL a VA did not give a confidence level. GPT-4V gave some more detailed information in the reason, such as the spire and steeple. On the contrary, LLaVA gave a partially incorrect description of the reason, such as boat and water.

GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['Alarm_Clock', 'Backpack', …, 'Webcam']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: Calendar
Confidence Score: 0.95
Reasoning: The image shows a **grid** that resembles a **monthly calendar layout**, with numbers that seem to represent **dates** at the top, which is a common feature of calendars.

LLaVA:

Answer Choice: Calendar
Confidence Score: 0.90
Reasoning: The image shows a **calendar** with a **clock** on it.

Figure 35: Induced Distribution Shift: Case 3 - Analyzing the 'dog' Category in the art_painting Domain of the PACS_unseen Dataset. In this case study, both GPT-4V and LLaVA models are tasked with responding to an identical text prompt accompanied by an image. Both GPT-4V and LLaVA predicted correctly, however, GPT-4V gave a confidence level of 1.00 and LLaVA did not give a confidence level. Both GPT-4V and LLaVA give a more nuanced and reliable reason.

GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['dog', 'elephant', …, 'person'].\\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: dog
Confidence Score: **1.00**
Reasoning: The image clearly shows a young **dog** with **distinct facial markings and fur** that is typical of a **canine**.

LLaVA:

Answer Choice: dog
Confidence Score: **None**
Reasoning: The image is of a **dog with blue and white fur**.

Figure 36: Natural Distribution Shift: Case 4 - Analyzing the 'Calendar' Category in the Clipart Domain of the Office-Home_unseen Dataset. In this case study, both GPT-4V and LLaVA models are tasked with responding to an identical text prompt accompanied by an image. Both GPT-4V and LLaVA predicted correctly, but GPT-4V gave higher confidence. GPT-4V and LLaVA focus on different details. GPT-4V focuses on the grid and dates associated with Calendar, whereas LLaVA focuses on another object 'clock' in the image.

GPT-4V: Correct | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [‘Alarm_Clock’, ‘Backpack’, …, ‘Webcam’]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: Postit_Notes
Confidence Score: 0.95
Reasoning: The image appears to show a collection of Post-it notes arranged in a creative way to resemble various objects like windows, doors, and a sign. **Each Post-it note has writing or drawing on it, which is characteristic of how Post-it notes are used for reminders, notes, or labels**. The overall arrangement is artistic, but the primary objects depicted are **clearly Post-it notes**.

LLaVA:

Answer Choice: Postit_Notes
Confidence Score: 0.8
Reasoning: The image is a **collection of post it notes with different messages on them**.

Figure 37: Natural Distribution Shift: Case 5 - Analyzing the ‘Postit_Notes’ Category in the Art Domain of the Office-Home_unseen Dataset. In this case study, both GPT-4V and LLaVA models are tasked with responding to an identical text prompt accompanied by an image. Both GPT-4V and LLaVA predicted correctly, but GPT-4V gave higher confidence with more detailed description. GPT-4V focuses on the writing and drawing associated with Postit_Notes, whereas LLaVA focuses on massages on it. This also demonstrates that GPT-4V and LLaVA have a certain degree of generalization ability on unseen data with domain shifts, with GPT-4V possessing stronger explanatory capabilities.

GPT-4V: Incorrect | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [‘Alarm_Clock’, ‘Backpack’, …, ‘Webcam’]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

Ground-truth Answer: TV

GPT-4V:

Answer Choice: Printer
Confidence Score: 0.95
Reasoning: The item in the top left section of the large shelving unit appears to **be a printer due to its shape, the presence of paper trays,** and its position atop what seems to be a stack of paper, which is a common accessory with printers. The overall configuration of the item, with what seems like control buttons and possibly a display, is characteristic of a printer's design.

LLaVA:

Answer Choice: Postit_Notes
Confidence Score: 0.8
Reasoning: The image shows a display of various items, including a **display of post it notes**.

Figure 38: Natural Distribution Shift: Case 6 - Analyzing the ‘TV’ Category in the Real World Domain of the Office-Home_unseen Dataset. This image presents a TV displaying complex content. Due to the misleading nature of the complex content, both GPT-4V and LLaVA made mistakes. GPT-4V misidentified the TV as a Printer, while LLaVA misidentified it as Post-it Notes. This result demonstrates that both GPT-4V and LLaVA still have difficulties in predicting complex samples accurately.

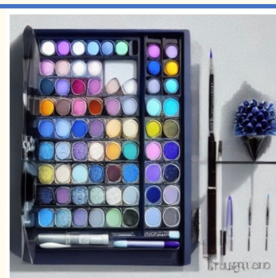
GPT-4V: Incorrect | LLaVA: Correct

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: ['Alarm_Clock', 'Backpack', …, 'Webcam']. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.



GPT-4V:

Answer Choice: Push_Pin
Confidence Score: 0.95
Reasoning: The image includes what appears to be two pens on the right-hand side, **next to the push pins.**

LLaVA:

Answer Choice: Marker
Confidence Score: 0.8
Reasoning: The image shows a variety of **art** supplies, including **a palette of colors and a paintbrush**. The presence of a palette and paintbrush suggests that the image is related to art, and **markers are commonly used in artistic projects.**

Figure 39: Natural Distribution Shift: Case 7 - Analyzing the 'Marker' Category in the Art Domain of the Office-Home_unseen Dataset. In this case, LLaVA correctly identifies the object as a 'Marker' with a confidence score of 0.8, whereas GPT-4V, with a confidence score of 0.95, mistakenly identifies the object as a 'Push_Pin'. Due to the presence of an object resembling a Push_Pin in the image, GPT-4V identified the image as a Push_Pin. Meanwhile, not only did LLaVA correctly predict, but it also provided a description related to its prediction: a palette of colors and a paintbrush.

GPT-4V: Correct | LLaVA: Incorrect

Prompt:

Given the image, answer the following question using the specified format. \\
Question: What is in this image? \\
Choices: [‘Alarm_Clock’, ‘Backpack’, …, ‘Webcam’]. \\

Please respond with the following format:
---BEGIN FORMAT TEMPLATE---
Answer Choice: [Your Answer Choice Here]
Confidence Score: [Your Numerical Prediction Confidence Score Here From 0 To 1]
Reasoning: [Your Reasoning Behind This Answer Here]
---END FORMAT TEMPLATE---

Do not deviate from the above format. Repeat the format template for the answer.

GPT-4V:

Answer Choice: Clipboards
Confidence Score: 1.0
Reasoning: The image shows **a blue clipboard with a metal clip at the top**, which is used to hold papers securely.

LLaVA:

Answer Choice: Postit_Notes
Confidence Score: 0.8
Reasoning: The image is a collage of various items, and one of the items is **a post it note.**

Figure 40: Natural Distribution Shift: Case 8 - Analyzing the ‘Clipboards’ Category in the Clipart Domain of the Office-Home_unseen Dataset. In this image, GPT-4V accurately identifies the object as Clipboards, noting a metal clip at the top of object, and assigning a confidence score of 1.0. GPT-4V successfully captured the key element ‘clip,’ which helped in identifying the object as Clipboards. In contrast, LLaVA incorrectly classifies the object as Postit_Notes with a confidence score of 0.8, failing to recognize the key element ‘clip’ of Clipboards.