On the Out-Of-Distribution Generalization of Large Multimodal Models

Xingxuan Zhang[1†], Jiansheng Li[1†], Wenjing Chu[2], junjia hai[2], Renzhe Xu[1], Yuqing Yang[1], Shikai Guan[1], Jiazheng Xu[1], Liping Jing[2], Peng Cui[1*]
[1]Tsinghua University, [2]Beijing Jiaotong University
xingxuanzhang@hotmail.com, cuip@tsinghua.edu.cn

Abstract

We investigate the generalization boundaries of current Large Multimodal Models (LMMs) under out-of-distribution scenarios and domain-specific tasks. We evaluate their zero-shot generalization across synthetic images, real-world distributional shifts, and specialized datasets like medical and molecular imagery. Empirical results indicate that LMMs struggle with generalization beyond common training domains, limiting their direct application without adaptation. To understand the cause of unreliable performance, we analyze three hypotheses: semantic misinterpretation, visual feature extraction insufficiency, and mapping deficiency. Results identify mapping deficiency as the primary hurdle. To address this problem, we show that in-context learning (ICL) can significantly enhance LMMs' generalization. We further explore the robustness of ICL under distribution shifts and show its vulnerability to domain shifts, label shifts, and spurious correlation shifts between in-context examples and test data, opening new avenues for overcoming generalization barriers.

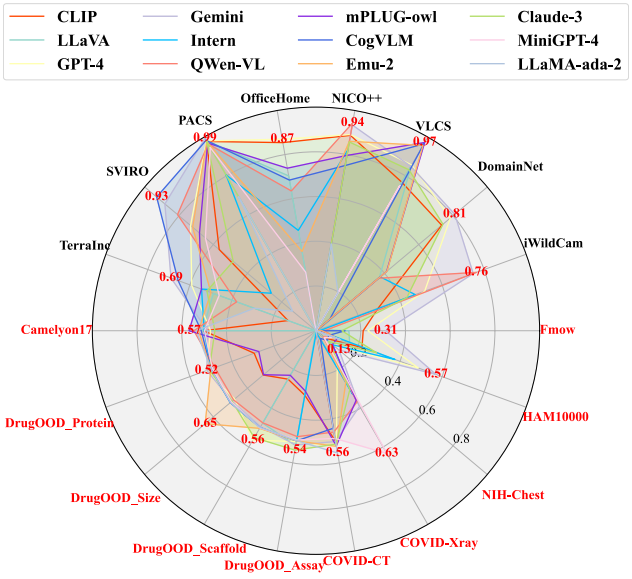

Figure 1. Zero-shot generalization of current LMMs on OOD generalization or domain-specific tasks.

1. Introduction

Supported by the exponential growth of both data availability and computational prowess, the field of large language models (LLMs) [22, 80] has witnessed a surge in interest and development, demonstrably achieving generalization across diverse natural language tasks [65, 110]. Pushing the boundaries of general artificial intelligence further, the rise of Large Multimodal Models (LMMs) [2, 26] aims to integrate LLM strengths with additional sensory modalities, such as image, audio, or 3D data, enabling them to condition their output on this broader range of inputs [79, 99]. Some works extend LLMs with visual understanding through end-to-end tuning, such as GPT-4 [94], Gemini [79], Flamingo [2], and Qwen [9]. Alternatively, fueled by powerful LLMs like LLaMA [80], modular combinations of LLMs and image-to-text models such as LLaVA [58], MiniGPT-4 [110], InstructBLIP [26] and many more [7, 99] are also being explored. The burgeoning field of LMMs has also witnessed a surge in benchmark development [10, 35, 59], which demonstrate LMMs' intriguing ability to solve complex multimodal tasks.

Despite the considerable progress in LMMs, a comprehensive understanding of their generalization capabilities remains elusive, particularly when applied to specialized domains or confronted with distributional shifts [96]. Recently, researchers have highlighted the potential impact of distributional bias on the reliability of LLMs [84, 96]. Notably, even GPT-4, a state-of-the-art LMM, exhibits susceptibility to generating erroneous outputs when encountering distribution shifts [40, 94]. Yet comprehensive evaluations of current LMMs under out-of-distribution scenarios and on domain-specific tasks rarely addressed during training are scarce. Furthermore, existing studies often lack

†Equal contribution, *Corresponding author
Code is available at https://github.com/jiansheng-li/OOD-Generalization-of-LMMs

in-depth analysis and explanation of the underlying causes driving LMMs' susceptibility to errors in these situations. This knowledge gap hinders the practical implementation of LMMs and necessitates further research efforts in diverse application areas under varying data distributions.

In this paper, we aim to delineate the boundaries of the current LMMs' generalization beyond the confines of its training data distribution. We initially assess the zero-shot generalization capabilities of LMMs across diverse test data including images under synthetically generated and natural distribution shifts, as well as domain-specific imagery. Our objective is to simulate real-world data variations and evaluate the applicability of these models to specialized tasks. We find that current LMMs struggle with zero-shot generalization beyond domains closely resembling their training data. Direct application of LMMs to specific domains without further adaptation or fine-tuning is likely to be insufficiently accurate and reliable.

To uncover the root cause of unreliability, we conduct a comprehensive error analysis investigating three plausible hypotheses: (1) Semantic misinterpretation: LMMs may struggle to grasp the nuanced meaning and implications of specific scientific categories within prompts, harming their understanding of the task. (2) Visual feature extraction insufficiency: The inherent characteristics of medical and molecular data, like high dimensionality or intricate image features, might challenge the model's internal encoding and processing mechanisms, hindering accurate visual information extraction. (3) Mapping deficiency: Restricted training data in specialized domains hinders the development of robust mappings between semantic meanings and visual features in LMMs.

Our analysis identifies that mapping deficiency is the primary hindrance to model generalization. To address this limitation, we investigate the potential of in-context learning (ICL) to enhance the model's acquisition and utilization of critical relationships between semantic descriptions and visual features. Surprisingly, we find that incorporating in-context examples (ICE) not only from the target distribution but also those with domain shifts significantly enhances the generalization capabilities of LMMs. This opens exciting avenues for using designed ICL to break through generalization barriers. Furthermore, we show that some severe distribution shifts between ICE and test samples, including label and spurious correlation shifts, can introduce performance decline and instability rising. Thus promising data selection of ICE approaches or reinforcement of generalization is required for reliable applications of LMMs on specific tasks.

The contribution of the paper is summarized as follows.

- We evaluate the zero-shot generalization of 15 current LMMs on 20 datasets under various distributional shifts. We demonstrate that the OOD generalization performance of LMMs can significantly diverge from their performance on current public generic benchmarks.
- Our analysis identifies that mapping deficiency other than semantic misinterpretation and visual feature extraction insufficiency can be the primary hindrance to model generalization.
- We validate the potential of ICL to enhance the model's mapping knowledge and boost the generalization of LMMs.
- We systematically examine the impact of distributional bias within ICE, revealing that label and spurious correlation shifts significantly hinder the generalization of LMMs, while domain shifts do not.

2. Zero-shot Generalization

This section delves into the zero-shot generalization performance of LMMs on visually shifted data encompassing three categories: synthetically generated images, naturally occurring distribution shifts, and domain-specific imagery like medical and molecular datasets. We aim to elucidate the generalization limitations and strengths of LMMs under distinct data variations. We consider LMMs including LLaVA [58], QWen [9], CogVLM [85], mPLUG-owl [95], MiniGPT-4 [110], LLaMA-adapter V2 [36], CLIP [71], BLIP [53], InstructBLIP [26], kosmos-2 [68], Emu-2 [77], Intern [98], Claude-3 [3], Gemini-1.5 [79] and GPT-4 [65][1] for our experiments. The primary results are presented in Figure 1, where datasets exhibiting superior LMMs performance are labeled in black, and those with inferior performance are labeled in red.

2.1. Zero-shot Generalization on Synthetic and Natural Images

We present an evaluation of LMMs on imagery with significant synthetic and natural shifts in data distribution. We leverage 11 datasets, each with its distinct characteristics and challenges: CMNIST [5], RMNIST [37], PACS [51], VLCS [33], OfficeHome [83], TerraInc [12], DomainNet [67], Fmow [23], iWildCam [13], SVIRO [25], and NICO++ [102]. The introduction of these datasets can be found in Appendix C.

LMMs achieve outstanding performance on most of the well-used datasets for domain generalization or OOD generalization, such as PACS, VLCS, OfficeHome, DomainNet, and NICO++. The performance approaches or even surpasses that of the existing optimal algorithms, suggesting that **the generalization of these models to recognize common objects in the domains contained in these datasets is strong.** This may be because these domains have been included in the training data for these large models. Notably, the usability of current common ob-

[1]Please refer to Appendix C for the details of versions of these LMMs.

Dataset	CMNIST	RMNIST	DomainNet	Fmow	iWildCam	NICO++	OfficeHome	PACS	SVIRO	TerraInc	VLCS	Average
LLaVA	0.622	0.696	0.480	0.148	0.054	0.849	0.736	0.980	0.874	0.668	**0.975**	0.644
QWen2-VL	0.325	0.403	0.755	**0.340**	0.475	0.906	0.750	0.981	0.802	0.585	0.949	0.526
CogVLM	0.117	0.317	0.081	0.013	0.000	0.066	0.684	0.983	0.930	0.656	0.965	0.437
mPLUG-owl	0.356	0.411	0.412	0.020	0.003	0.796	0.738	0.969	0.680	0.545	0.972	0.537
MiniGPT-4	0.285	0.294	0.000	0.062	0.000	0.631	0.266	0.945	0.645	0.461	0.750	0.394
LLaMA-ada	0.314	0.252	0.389	0.022	0.003	0.394	0.000	0.690	0.137	0.294	0.197	0.245
CLIP	0.468	0.446	0.767	0.211	0.296	0.887	0.854	0.977	0.565	0.133	0.734	0.576
BLIP-2	0.339	0.376	0.508	0.043	0.000	0.673	0.527	**1.000**	**0.918**	0.328	0.937	0.514
InstructBLIP	0.510	0.611	0.477	0.020	0.000	0.483	0.003	0.723	0.000	0.376	0.257	0.315
kosmos-2	0.000	0.000	0.227	0.035	0.000	0.163	0.000	0.792	0.413	0.365	0.569	0.233
Emu-2	0.622	0.668	0.415	0.027	0.000	0.860	0.360	0.977	0.720	0.500	0.957	0.555
Intern	0.198	0.285	0.375	0.015	0.000	0.472	0.456	0.805	0.261	0.545	0.822	0.385
Gemini-1.5	**0.880**	**0.955**	**0.810**	0.312	**0.748**	**0.934**	**0.923**	0.993	0.883	**0.690**	0.849	**0.816**
GPT-4	0.772	0.895	0.784	0.232	0.517	0.891	0.870	0.980	0.732	0.591	0.531	0.709
Claude3	0.704	0.654	0.741	0.116	0.432	0.857	0.795	0.979	0.483	0.519	0.831	0.646
Random	0.100	0.100	0.003	0.116	0.091	0.017	0.015	0.143	0.333	0.143	0.200	0.115
SOTA	0.721[31]	0.981[76]	0.655[21]	0.380[19]	0.343[58]	0.885[102]	0.906[48]	0.990[55]	0.920[25]	0.696 [105]	0.855[62]	0.757

Table 1. Zero-shot generalization results of current state-of-the-art LMMs on OOD generalization datasets. *Random* indicates random guess and *SOTA* indicates previous SOTA on each dataset for comparison.

ject recognition is also evidenced by the exceptional performance of GPT-4 and Gemini on challenging benchmarks like DomainNet (74.8% and 75.3% accuracy, respectively) and iWildCam (52.3% and 68.2%). This surpasses the previous state-of-the-art approaches [21, 74]. The superior performance of GPT-4 and Gemini on benchmarks with numerous categories (DomainNet: 345, iWildCam: 206) suggests these models excel at handling complex prompts.

2.2. Zero-shot Generalization on Domain-Specific Data

We consider medical and molecular images for the evaluation of domain-specific generalization ability for the following reasons. 1) Analyzing LMMs' performance across these two disparate domains, with their unique knowledge requirements and data distributions, provides insights into the generalization capabilities and potential of LMMs for broader real-world applications. 2) They help investigate the potential of incorporating auxiliary inputs, such as domain-specific information and in-context examples, for enhancing the model's performance in specific application areas. 3) Failure analysis on these tasks can highlight areas where further development and domain-specific adaptation might be necessary. We consider diverse medical and molecular image datasets including Camelyon17 [78], CT-XCOV [32], XCOVFour [39], HAM10000 [81], NIH-Chest [64], and DrugOOD [44].

While LMMs excel at zero-shot generalization on OOD datasets of synthetic and natural images, their performance plummets when transitioning to medical and molecular data. Specifically, all LMMs make near-random predictions on all medical and molecular datasets. Please note that the average number of categories in these domain-specific datasets is significantly lower than that of common object datasets in Section 2.1, e.g., there are only 2 categories in Camelyon17, CT-XCOV, DrugOOD_Assay, DrugOOD_Scaffold, DrugOOD_Size, and DrugOOD_Protein.

This significant drop in performance suggests the crucial fact that LMMs struggle with zero-shot generalization beyond domains closely resembling their training data. Thus direct application of these models to medical and molecular domains without further adaptation or fine-tuning is insufficiently accurate and reliable. We study the underlying causes of the observed limitations of LMMs generalization on specific domains and potential solutions to enhance LMMs adaptability and robust performance across diverse domains in Section 3 and 4.

3. Failure Analysis

We consider three potential causes of the failure on domain-specific data: (1) Semantic misinterpretation: LMMs may struggle to grasp the nuanced meaning and implications of specific scientific concepts within prompts, leading to semantic misinterpretation within text inputs. (2) Visual feature extraction insufficiency: The inherent characteristics of medical and molecular data, such as high dimensionality and intricate image features, can challenge the model's encoding and processing mechanisms, hampering accurate visual information extraction. (3) Mapping deficiency: limited training data in specialized domains restricts the development of robust and generalizable mappings between semantic meanings and visual features.

Potential shared biases among LMMs. To uncover potential shared biases arising from model architecture and training methodologies in various LMMs, we begin with the analysis of individual error patterns across diverse datasets. We present an overall analysis of error in Appendix D and extensive case studies of zero-shot generalization of LMMs in Appendix E. We reveal a lack of consistent, dataset-specific biases. Instead, errors appeared scattered across most datasets. This suggests that **despite similar architectural and training characteristics, different LMMs do**

Dataset	Camelyon17	HAM10000	NIH-Chest	XCOVFour	CT-XCOV	DrugOOD_A	DrugOOD_Sc	DrugOOD_Si	DrugOOD_P	Average
LLaVA	0.508	0.363	0.089	0.431	0.509	**0.623**	**0.624**	0.623	**0.620**	**0.488**
QWen2-VL	**0.588**	0.185	0.044	0.591	0.450	0.577	0.405	0.433	0.498	0.419
CogVLM	0.513	0.061	0.066	0.040	0.443	0.500	0.500	0.500	0.500	0.347
MiniGPT-4	0.499	0.090	0.048	**0.633**	0.489	0.496	0.502	0.495	0.493	0.416
LLaMA-ada	0.536	0.117	0.000	0.000	0.555	0.499	0.500	0.500	0.500	0.356
CLIP	0.502	0.219	0.050	0.000	0.500	0.292	0.250	0.308	0.294	0.268
BLIP-2	0.503	0.073	0.006	0.333	0.503	0.500	0.500	0.500	0.500	0.380
kosmos-2	0.189	0.012	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.022
Emu-2	0.525	0.103	0.047	0.267	0.525	0.500	0.504	**0.647**	0.500	0.402
Intern	0.500	0.376	0.098	0.328	0.008	0.500	0.500	0.500	0.500	0.368
Gemini-1.5	0.485	**0.571**	**0.134**	0.390	0.503	0.519	0.495	0.491	0.519	0.456
GPT-4	0.472	0.487	0.098	0.190	**0.526**	0.504	0.565	0.475	0.521	0.426
Claude3	0.452	0.288	0.075	0.303	0.515	0.542	0.552	0.490	0.508	0.414
Random	0.500	0.143	0.050	0.250	0.500	0.500	0.500	0.500	0.500	0.383
SOTA	0.948[43]	0.965[47]	0.294[41]	0.955[14]	0.930[32]	0.821[44]	0.763[44]	0.714[44]	0.804[44]	0.799

Table 2. Zero-shot generalization results of LMMs on domain-specific datasets.

not exhibit consistent biases in semantic or visual feature extraction, such as persistently associating specific semantic features with incorrect visual features. Moreover, this observation raises the intriguing possibility of constructing multimodal models by strategically combining LMMs to leverage their diverse error profiles and achieve enhanced generalization capabilities.

Semantic misinterpretation. To analyze the impact of semantic misinterpretation within text input on model failure, we use informative and distinct prompts that provide additional scientific context and guidance for each dataset. These prompts aim to address the potential limitations of LMMs in understanding specific meanings and implications of certain categories within the input prompt, particularly in specialized scientific contexts. Specifically, we crafted introductory prompts for each dataset comprising three key elements: (1) domain specification and context, outlining the relevant field and study goals; (2) expected expertise, guiding the model towards domain-specific knowledge; and (3) category-specific introductions, providing clear explanations of each category's meaning and potential characteristics. Please see Appendix D.2 for details.

Table 3 shows that augmenting prompts with detailed domain-specific background information yields negligible performance improvements for LMMs in these tasks. This suggests that: (1) semantic misinterpretation within text input might not be the main limiting factor, or (2) the model's internal mechanisms may be inadequate for efficiently integrating and leveraging the auxiliary information provided.

To thoroughly verify whether the models truly comprehend questions and the supplementary information provided in prompts, we conduct a model semantic understanding experiment. We task GPT-4o with formulating three questions based on a standard prompt text (as shown in Appendix C.4) and ten questions based on an informative and contextual prompt (as shown in Appendix D.2). Subsequently, we prompt other LMMs to answer these questions based on their corresponding prompts. GPT-4o then evaluates their responses, assessing their comprehension of the prompt's textual content and assigning scores from 0 to 10 accordingly. Note that to maintain consistency with previous experimental setups, we include images in the prompts for this experiment as well, thus eliminating any potential influence of image inclusion on the results. To establish baselines for GPT-4o's scoring, we evaluate the scores of GPT-2 and randomly generated responses. These random responses are generated by having LMMs answer the questions posed by GPT-4o and then shuffling their answers randomly before presenting them to GPT-4o for scoring. Details are in Appendix D.3. As shown in Table 4, compared to the baseline, LMMs demonstrate significantly higher semantic understanding capabilities, fundamentally grasping the task within the prompt and providing reasonable answers.

Furthermore, we examine the potential influence of category order within the prompt, aiming to ensure equitable impact on model output regardless of category position. Analysis presented in the Appendix D.4 indicates that category order does not significantly affect model performance. **Therefore, semantic understanding failure is unlikely to severely affect LMMs generalization.**

Visual feature extraction insufficiency. To analyze the impact of data complexity within visual input, we investigate the sufficiency of visual extraction capabilities in LMMs for generalizing to specific domains. We employ a linear probing approach, utilizing the publicly available CLIP model as a feature extractor. Extracted features from CLIP are then fed into a separate linear classifier trained with corresponding category labels. By analyzing the classifier's accuracy on unseen test data, we assess the quality of visual features extracted by LMMs for generalization.

Table 5 demonstrates that when tested using linear probing, the feature extractors of open-source LMMs exhibit comparable performance to CLIP, and all significantly outperform the zero-shot approach on medical and molecule data. Thus the visual features extracted by LMMs and CLIP possess the capability to distinguish categories in these

Dataset	Camelyon17	HAM10000	NIH-Chest	XCOVFour	CT-XCOV	DrugOOD_A	DrugOOD_Sc	DrugOOD_Si	DrugOOD_P	Average
LLaVa	0.555(+4.7%)	0.621(+25.8%)	0.097(+0.8%)	0.327(-10.4%)	0.333(-17.6%)	0.501(-12.2%)	0.529(-9.5%)	0.504(-11.9%)	0.512(-10.8%)	0.442(-4.6%)
QWen-VL	0.523(-2.3%)	0.101(+0.6%)	0.078(+3.6%)	0.329(-3.8%)	0.500(+4.0%)	0.547(+6.6%)	0.502(+2.6%)	0.538(+5.7%)	0.504(+2.3%)	0.402(+2.1%)
CogVLM	0.500(-1.3%)	0.179(+11.8%)	0.000(-6.6%)	0.100(+6.0%)	0.500(+5.7%)	0.500(+0.0%)	0.001(-49.9%)	0.647(+14.7%)	0.629(+12.9%)	0.340(-0.7%)
MiniGPT-4	0.502(+0.3%)	0.223(+13.3%)	0.000(-4.8%)	0.008(-62.5%)	0.486(-0.3%)	0.000(-49.6%)	0.000(-50.2%)	0.000(-49.5%)	0.000(-49.3%)	0.135(-28.1%)
LLaMa-ada	0.439(-9.7%)	0.004(-11.3%)	0.000(+0.0%)	0.022(+2.2%)	0.441(+44.1%)	0.500(-5.5%)	0.500(+0.1%)	0.500(+0.0%)	0.500(+0.0%)	0.363(+0.7%)
BLIP-2	0.500(-0.3%)	0.004(-6.9%)	0.010(+4.0%)	0.333(+0.0%)	0.500(-0.3%)	0.500(+0.0%)	0.500(+0.0%)	0.500(+0.0%)	0.500(+0.0%)	0.372(-0.8%)
Emu-2	0.575(+5.0%)	0.098(-0.5%)	0.000(-4.7%)	0.333(+6.6%)	0.500(-2.5%)	0.717(+21.7%)	0.473(-3.1%)	0.504(-14.3%)	0.512(+1.2%)	0.489(+8.7%)
Intern	0.500(+0.0%)	0.000(-37.6%)	0.087(-1.1%)	0.205(-12.3%)	0.500(+0.0%)	0.500(+0.0%)	0.496(-0.4%)	0.500(+0.0%)	0.500(+0.0%)	0.365(-0.8%)
Gemini	0.506(+2.1%)	0.410(-16.1%)	0.104(-3.0%)	0.427(+3.7%)	0.697(+19.4%)	0.534(+1.5%)	0.506(+1.1%)	0.528(+3.7%)	0.519(+0.0%)	0.470(+1.4%)
GPT-4	0.430(-4.2%)	0.291(-19.0%)	0.000(-9.8%)	0.100(-9.0%)	0.534(+0.8%)	0.050(-45.4%)	0.023(-54.2%)	0.060(-41.5%)	0.295(-22.6%)	0.198(-22.8%)

Table 3. Prompt contextual information results on domain-specific datasets. Performance differences versus the common prompt are indicated parenthetically, with positive values in red reflecting improvements in accuracy and negative values in green indicating drops.

	Camelyon17_base	Camelyon17_aug	XcovFour_base	XcovFour_aug	DrugOOD_base	DrugOOD_aug
Gemini-1.5_random	1.33±0.58	2.88±2.94	2.00±1.00	2.33±2.75	2.00±0.00	1.90±0.74
Gemini-1.5	9.33±1.15 (+8.00)	9.80±0.42 (+6.92)	10.00±0.00 (+8.00)	9.75±0.44 (+7.42)	9.67±0.58 (+7.67)	9.10±0.74 (+7.90)
QWen2-VL_random	2.00±1.00	3.00±2.87	1.33±0.58	2.70±1.25	2.00±1.00	2.40±1.84
QWen2-VL	9.67±0.58 (+7.67)	9.75±0.44 (+6.75)	10.00±0.00 (+8.67)	9.20±0.92 (+6.50)	9.60±0.70 (+7.60)	9.00±0.94 (+6.60)
GPT-2	5.67±1.15	5.33±0.58	6.67±1.15	3.30±1.18	6.33±1.53	4.50±2.22

Table 4. Scores for LMMs' understanding of questions and informative prompts. *_base* denotes the use of base prompts while *_aug* refers to prompts with extra information.

tasks, and effective classification can be achieved when coupled with a suitable linear classifier. **This suggests that the visual feature extraction is unlikely the bottleneck hindering LMMs generalization to these specific tasks.**

The scaling of zero-shot generalization. We explore the potential of enhanced input encoding capabilities in large language models to quantify and potentially mitigate the limitations in generalization arising from insufficient encoding capacity of both text and visual inputs. Extensive research has established the performance gains achieved through scaling neural networks, often governed by reliable scaling laws related to training data, model size, and computational resources[20, 65]. This understanding proves invaluable in mitigating the escalating cost of large-scale experiments. However, existing scaling laws primarily apply to in-distribution performance or general tasks, neglecting OOD generalization. We investigate the scaling laws for LMMs under OOD scenarios, shedding light on their behavior in real-world contexts.

Due to the limitations in publicly available model sizes for other LMMs, we use CLIP as a representative model to investigate the scalability of zero-shot generalization. Specifically, we explore CLIP models with different sizes of ViT architecture to evaluate the performance of these CLIP models across five datasets: DomainNet, Terra Incognita, Office-Home, NICO, and SVIRO, which are suited for showcasing effective CLIP performance while minimizing potential data leakage biases, thus maximizing the reliability of our conclusions regarding the effectiveness of scaling.

Unlike the typical scaling law observed on in-distribution tasks, where larger models tend to perform better, Figure 2 reveals a different trend for OOD tasks, even exhibiting a notable decline on TerraInc and SVIRO. **This suggests that simply scaling up model size and improving its in-domain encoding skills may not be sufficient to guarantee improved generalization ability in domain-specific scenarios.**

We then analyze the impact of mapping deficiency and potential solutions with in-context learning in Section 4.

4. In-context Learning Generalization

Building upon the failure analysis in Section 3, this section investigates the potential of in-context learning (ICL) to bridge the gap between LMMs and domains not represented in their pre-training and fine-tuning data by addressing the mapping deficiency. ICL has recently garnered significant attention as a potential paradigm shift for large vision models [11, 107]. The following sub-sections study the potential of ICL for enhancing LMMs generalization by exploring two avenues: 1) Utilizing target domain data directly to inform the selection and construction of in-context examples (ICE). 2) Employing ICE that differ from the target domain data. This focuses on evaluating the effectiveness of ICL when bridging the gap between known and target data distributions. In other words, we aim to assess LMMs' performance when the distribution of test data is not precisely known when designing ICE.

Notations. Let X denote the inputs and Y the outputs. Suppose the test distribution is $P^{\text{test}}(X, Y)$ and ICE $\{(x_i, y_i)\}_{i=1}^{n}$ are sampled from a distribution $P^{\text{ICL}}(X, Y)$.

4.1. ICL with Data from Target Distribution

This subsection explores the effectiveness of ICL to LMMs in an ideal scenario: when examples are sampled directly from the target distribution, i.e., $P^{\text{ICL}}(X, Y) = P^{\text{test}}(X, Y)$. This simulates a setting where the target domain's data distribution is known and readily available for crafting ICE. By investigating the model's performance under such optimal conditions, we aim to determine the up-

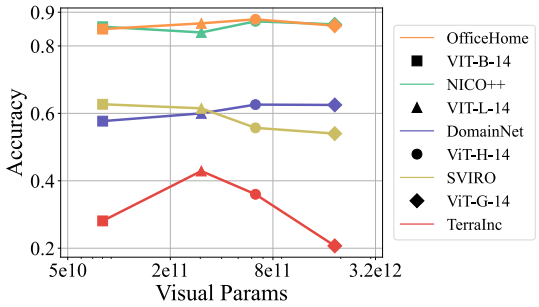

Figure 2. Scaling law of zero-shot generalization under distribution shifts.

Table 5. The comparison of CLIP with linear probing (donated as CLIP(LP)) and zero-shot generalization of CLIP, GPT-4, and Gemini.

	CLIP (LP)	CLIP	GPT-4	Gemini
COVID-CT	0.830	0.500 (-33 .0%)	0.432 (-47.95%)	0.498 (-40.00%)
DrugOOD_Assay	0.760	0.292 (-61.58%)	0.421 (-44.61%)	0.505 (-33.55%)
DrugOOD_Size	0.771	0.308 (-60.05%)	0.432 (-43.97%)	0.499 (-35.28%)
Fmow	1.000	0.211 (-78.90%)	0.220 (-78.00%)	0.268 (-73.20%)
HAM10000	0.840	0.219 (-73.93%)	0.530 (-36.90%)	0.410 (-51.19%)
NIH-Chest	0.740	0.050 (-93.24%)	0.057 (-92.30%)	0.078 (-89.46%)
COVID-Xray	0.970	0.000 (-97.00%)	0.239 (-75.36%)	0.390 (-59.79%)

per bound of potential performance improvements achievable through ICL. This theoretical understanding can inform real-world applications by illuminating the limitations and optimal capabilities of this approach when adapting LMMs to specific domains.

Given the previously observed underperformance of LMMs on specific domain-specific datasets, we select challenging and diverse datasets for our ICL experiments. This choice prioritizes datasets where LMMs have demonstrated limitations, allowing us to investigate the potential of ICL to bridge this performance gap. We analyze GPT-4 and Gemini due to their superior ICL capabilities observed in our preliminary experiments, which enables the isolation and analysis of the effectiveness of ICL without confounding factors arising from models' inherent limitations in understanding the guidance provided within such examples.

To further control for potential confounding factors like prompt understanding and category imbalance, we implemented a simplified experimental design for evaluating ICL's impact on LMMs performance on new domains and tasks. Across all datasets, we first randomly partitioned both domain and category into two balanced groups for the test data, ensuring an equal number of samples per category. This mitigates the influence of category imbalance and potentially improves the consistency and reliability of our conclusions. We then investigate the effectiveness of ICE by systematically varying their number. This stratified sampling approach ensures that the ICE can be considered independently and identically distributed (i.i.d.), statistically resembling the expected test data distribution.

Both GPT-4 and Gemini exhibit substantial performance gains in ICL across XCOVFour, iWildCam, NIH-Chest, and HAM10000. This trend consistently held, with increasing numbers of ICE correlating with improved model performance, as shown in Figure 3. **The results demonstrate that when ICE closely mirror the distribution of test data, LMMs can achieve substantial performance gains on many new tasks, where the degree of improvement scales with the number of examples provided. This further suggests that the primary obstacle to generalization in these tasks likely stems from deficiencies in task-specific knowledge mapping within the LMMs.**

We show the whole results of GPT-4 and Gemini after ICL with data from target distributions in Table 6. Intriguingly, both GPT-4 and Gemini fail to achieve performance gains on molecular datasets. This may be due to the inherent complexity of predicting molecular activity. Accurately classifying a molecule as "active" or not often requires extensive domain knowledge and substantial training data, which mere ICE might not adequately provide. This observation suggests that ICL may not be a universally effective approach, particularly for tasks demanding specialized knowledge beyond readily learnable patterns. Please see Appendix D.6 for more discussion. Further investigation is needed to elucidate the specific factors hindering its efficacy in such cases and potentially refine ICL strategies for tasks requiring domain expertise or introduce other methods to improve the adaptation of models such as retrieval-augmented generation (RAG) [49].

4.2. ICL under Distribution Shifts

This subsection investigates **the generalization of LMMs with ICL under realistic conditions where the target data distribution may be unknown or deviate from that of the ICE**. We focus on three key challenges: (1) Domain shifts: ICE and test samples are drawn from distinct domains defined by each dataset. (2) Label shifts: While the underlying domain remains, the distribution of labels differs between ICE and test data. (3) Spurious correlation shifts: If ICE contain misleading correlations not present in the target domain, the model may be prone to biased decisions, indicating the careful design of ICE is needed to avoid such pitfalls. Please note that these three types of shifts are not orthogonal to each other, but they all represent the common and effective utilization of LMMs in real-world applications. While real-world domain shifts may potentially encompass both label and spurious correlations shifts, our experiments confine such shifts within the predefined domains of the dataset to isolate and quantify the individual effects of each factor on model generalizability.

	ICE	XCOVFour	iWildCam	NIH-Chest	CT-XCOV	DrugOOD_P	DrugOOD_Si	DrugOOD_A	Camelyon17	HAM10000
GPT-4	0	0.453(+0.0%)	0.642(+0.0%)	0.634(+0.0%)	0.500(+0.0%)	0.463(+0.0%)	0.303(+0.0%)	0.447(+0.0%)	0.500(+0.0%)	0.544(+0.0%)
	2	0.688(+23.5%)	0.728(+8.6%)	0.634(+0.0%)	0.471(-2.9%)	0.449(-1.4%)	0.325(+2.2%)	0.454(+0.7%)	0.500(+0.0%)	0.645(+10.1%)
	4	0.831(+37.8%)	0.890(+24.8%)	0.720(+8.6%)	0.658(+15.8%)	0.475(+1.2%)	0.292(-10.2%)	0.500(+5.3%)	0.568(+6.8%)	0.667(+12.3%)
	8	0.843(+39.0%)	0.927(+28.5%)	0.736(+10.2%)	0.591(+9.1%)	0.475(+1.2%)	0.355(+5.2%)	0.395(-5.3%)	0.452(-4.8%)	0.640(+9.6/%)
Gemini	0	0.742(+0.0%)	0.975(+0.0%)	0.585(+0.0%)	0.543(+0.0%)	0.48(+0.0%)	0.452(+0.0%)	0.488(+0.0%)	0.532(+0.0%)	0.533(+0.0%)
	2	0.870(+12.8%)	0.967(-0.8%)	0.626(+4.1%)	0.571(+2.8%)	0.532(+5.2%)	0.408(-4.4%)	0.465(-2.3%)	0.505(-2.7%)	0.708(+17.5%)
	4	0.842(+10.0%)	0.978(+0.3%)	0.663(+7.8%)	0.615(+7.2%)	0.563(+8.3%)	0.455(+0.3%)	0.438(-5.0%)	0.585(+5.3%)	0.656(+12.3%)
	8	0.798(+5.6%)	0.978(+0.3%)	0.693(+10.8%)	0.813(+27.0%)	0.546(+6.6%)	0.501(+4.9%)	0.465(-2.3%)	0.504(-2.8%)	0.756(+22.3%)

Table 6. Results of ICL with data from target distribution.

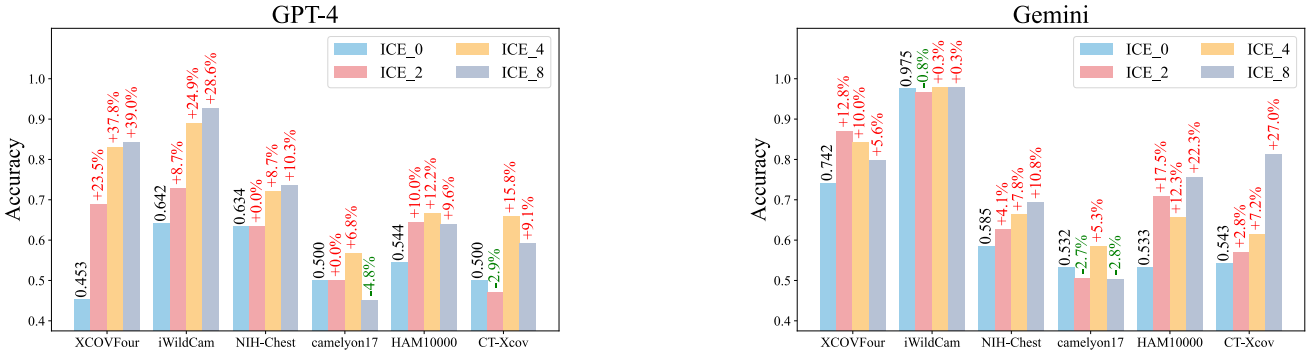

Figure 3. ICL Generalization of GPT-4 and Gemini with ICE from the target distribution.

4.2.1. ICL under Domain Shifts

We leverage the inherent domain diversity within existing datasets to investigate ICL's efficacy under realistic conditions. Prior research has established that domain shifts pose significant challenges for generalization [38]. Therefore, we deliberately construct scenarios where ICE and test samples are drawn from different domains within these datasets, replicating natural or artificially induced distribution shifts, i.e., $P^{\text{ICL}}(X, Y)$ and $P^{\text{test}}(X, Y)$ are the distributions of different domains predefined in each dataset, respectively. This is the basic setting for domain generalization (DG) and out-of-distribution (OOD) generalization, which enables us to observe how effectively ICL can bridge these gaps and enhance LMMs generalization to unknown domains. In the following experiments, we sample ICE and test samples from 2 domains within each dataset. For datasets with multiple domains, we randomly sample 2 of them with adequate diversity of categories and sufficient samples.

As shown in Figure 4, ICE with domain shifts consistently significantly outperform zero-shot performance and show surprisingly robust performance, only marginally falling short of the gains achieved from examples directly drawn from the target distribution. This observation highlights the remarkable OOD generalization capabilities of LMMs with ICL in traditional OOD settings when presented with non-ideal examples. Specifically, across datasets including iWildCam, DomainNet, CT-XCOV, and HAM10000, ICL under domain shifts significantly bolsters both GPT-4 and Gemini's performance. We also observe a consistent trend of increasing accuracy with the number of provided examples. **While ICL with domain shifts slightly underperforms compared to target-distribution examples, the performance remains indistinguishable and suggests a remarkable capacity of LMMs to generalize across domains via ICL.**

4.2.2. ICL under Label Shifts

We study another common challenge for generalization: label shift [8], where the ground-truth output (labels) between ICE and the test data exhibit distributional differences. Specifically, this means that $P^{\text{ICL}}(Y) \neq P^{\text{test}}(Y)$ while $P^{\text{ICL}}(X|Y) = P^{\text{test}}(X|Y)$. Unlike domain shifts, which alter the entire underlying data structure, label shifts primarily affect the output label space. Here we investigate whether the model's output preference can be influenced by ICE, leading to potentially misleading predictions. Specifically, both ICE and test samples are sampled from the same domain of each dataset with different distributions on Y.

Results are shown in Table 7. Changing the proportion of categories in the context demonstrably alters the predicted category proportions, subsequently influencing overall performance. Maintaining a fixed test data distribution while varying the in-context categories can lead to declining performance and rising instability. **This suggests that label shifts in real-world applications can significantly impact LMMs generalizability. Therefore, the judicious selection of ICE considering both the task specifics and the desired target data distribution is crucial for stable model performance.**

4.2.3. ICL under Spurious Correlation Shifts

We explore the impact of spurious correlations on ICL's efficacy. Spurious correlations arise when training data

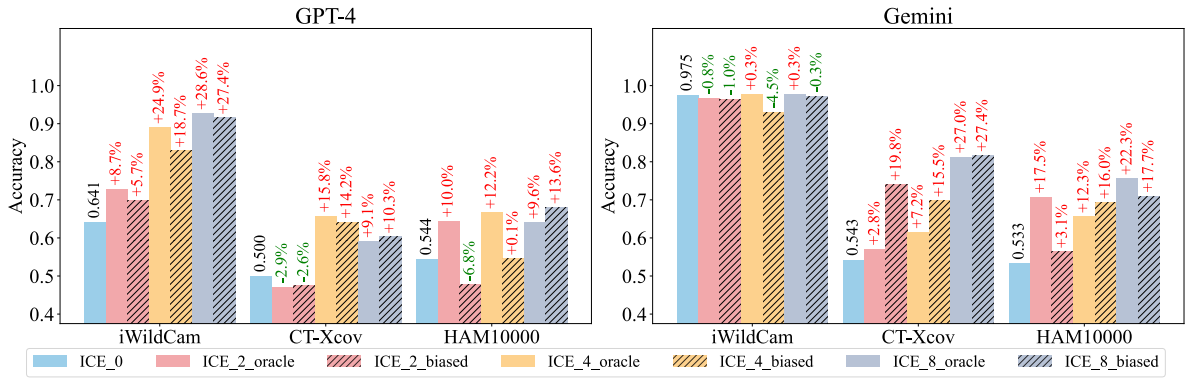

Figure 4. Generalization of GPT-4 and Gemini with ICL under domain shifts. ICE_2_oracle indicates ICL with 2 ICE from target distribution, while ICE_2_biased indicates 2 ICE that have domain shifts with target distribution.

	Ratio	XCOVFour	NIH-Chest	CT-XCOV
GPT-4	1:7	0.908(-0.9%)	0.719(-3.8%)	0.445(-9.5%)
	2:6	0.820(-9.7%)	0.707(-5.0%)	0.471(-6.9%)
	4:4	0.917(+0.0%)	0.757(+0.0%)	0.540(+0.0%)
	6:2	0.958(+4.1%)	0.675(-8.2%)	0.502(-3.8%)
	7:1	0.950(+3.3%)	0.663(-9.4%)	0.486(-5.4%)
Gemini	7:1	0.699(-12.9%)	0.603(-7.7%)	0.742(-4.8%)
	6:2	0.699(-12.9%)	0.627(-5.3%)	0.732(-5.8%)
	4:4	0.828(+0.0%)	0.680(+0.0%)	0.790(+0.0%)
	2:6	0.745(-8.3%)	0.676(-0.4%)	0.659(-13.1%)
	1:7	0.732(-9.6%)	0.641(-3.9%)	0.717(-7.3%)

Table 7. ICL Generalization under label shifts.

	ICE	CMNIST	DomainNet	iWildCam	Camelyon17	HAM10000
GPT-4	0	0.896	0.850	0.800	0.500	0.490
	2	0.490	0.730	0.860	0.530	0.530
	4	0.784	0.930	0.810	0.550	0.520
	8	0.803	0.970	0.880	0.470	0.510
Gemini	0	0.469	0.939	0.929	0.510	0.587
	2	0.641	0.919	0.786	0.520	0.626
	4	0.563	0.940	0.854	0.521	0.643
	8	0.583	0.860	0.936	0.541	0.646

Table 8. Results under spurious correlation. ICE indicates the number of ICE.

across domains exhibits statistically unreliable associations between input features and labels. Specifically, we divide the features X into invariant X_v and spurious X_s components, as outlined in Arjovsky et al. [5], Lin et al. [56]. The output Y depends solely on the invariant features X_v in both P^{ICL} and P^{test} distributions, represented as $Y = g(X_v, \epsilon)$, with $g(X_v, \epsilon)$ being the labeling function on X_v and ϵ as exogenous noise. The spurious correlation shift is characterized by the difference $P^{\text{ICL}}(Y|X_s) \neq P^{\text{test}}(Y|X_s)$.

These misleading correlations between Y and X_s, often superficially stronger than invariant dependencies between Y and X_v, can be readily acquired by models, hindering their ability to generalize beyond the training data's specific domain context. Specifically, we leverage datasets containing two distinct domains (D_1 and D_2) and two categories (C_1 and C_2). For ICE, we deliberately craft examples with contrasting domain-category combinations: C_1 data from D_1 and C_2 data from D_2. Conversely, test samples mirror the opposite scenario, presenting C_2 data from D_1 and C_1 data from D_2. This controlled setting ensures that ICE and test samples exhibit opposing domain-category associations, allowing us to isolate the influence of spurious correlation shifts between ICE and test samples on LMMs generalization beyond other distribution shifts.

The results in Table 8 support that spurious correlations can negatively impact the effectiveness of ICL. GPT-4 exhibits a marked drop in prediction accuracy across both CMNIST and DomainNet when subject to ICL. Similar observations are evident in Gemini's experiments on iWildCam and DomainNet. **This indicates that although ICL can contribute to generalization compared with zero-shot prediction in some scenarios, the presence of significant spurious correlations within ICE can pose a substantial threat to the generalization of LMMs on domains not included in their pre-training data. Thus promising data selection approaches or reinforcement of generalization against spurious correlation is required for reliable applications of LMMs on these tasks.**

5. Conclusion

We investigated the generalization of current LMMs and found that they showed unreliable performance beyond common training domains. We found that the mapping deficiency could be the major hurdle and showed the potential of ICL to largely enhance LMMs' generalization. We further presented ICL's vulnerability to label and spurious correlation shifts.

Acknowledgement

This work was supported in part by China National Postdoctoral Program for Innovative Talents (BX20240203), Beijing Municipal Science and Technology Project (No. Z241100004224009), NSFC(No. 62425206, 62141607).

References

[1] Kartik Ahuja and David Lopez-Paz. A closer look at in-context learning under distribution shifts, 2023. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, and et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[3] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 2024. 2

[4] Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *ArXiv*, abs/1907.02893, 2019. 1, 2

[5] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2, 8

[6] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *ArXiv*, abs/2110.10832, 2021. 1

[7] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, and et al. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023. 1

[8] Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019. 7

[9] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, and et al. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2

[10] Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*, 2023. 1

[11] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 5

[12] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 2

[13] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 2

[14] Pratik Bhowal, Subhankar Sen, Jin Hee Yoon, Zong Woo Geem, and Ram Sarkar. Choquet integral and coalition game-based ensemble of deep learning models for covid-19 screening from chest x-ray images. *IEEE Journal of Biomedical and Health Informatics*, 25(12):4328–4339, 2021. 4

[15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and et al. Language models are few-shot learners, 2020. 1

[16] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2224–2233, 2019. 1

[17] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, and et al. Swad: Domain generalization by seeking flat minima. In *Neural Information Processing Systems*, 2021. 1

[18] Stephanie C. Y. Chan, Ishita Dasgupta, Junkyung Kim, Dharshan Kumaran, Andrew K. Lampinen, and Felix Hill. Transformers generalize differently from information stored in context vs in weights, 2022. 1

[19] Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Somesh Jha, and Tomas Pfister. Aspest: Bridging the gap between active learning and selective prediction. *arXiv preprint arXiv:2304.03870*, 2023. 3

[20] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 5

[21] Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, and Suha Kwak. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15702–15712, 2023. 3

[22] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, and et al. Palm: Scaling language modeling with pathways, 2022. 1

[23] 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*, pages 6172–6180, 2018. 2

[24] Xu Chu, Yujie Jin, Wenwu Zhu, Yasha Wang, Xin Wang, Shanghang Zhang, and Hong Mei. Dna: Domain generalization with diversified neural averaging. In *International Conference on Machine Learning*, 2022. 1

[25] Steve Dias Da Cruz, Oliver Wasenmuller, Hans-Peter Beise, Thomas Stifter, and Didier Stricker. Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 973–982, 2020. 2, 3

[26] W Dai, J Li, D Li, AMH Tiong, J Zhao, and et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023. 1, 2

[27] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 2

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[29] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *Neural Information Processing Systems*, 2019. 1

[30] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, and et al. Palm-e: An embodied multimodal language model, 2023. 1

[31] Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022. 3

[32] Ismail Elbouknify, Afaf Bouhoute, Khalid Fardousse, Ismail Berrada, and Abdelmajid Badri. Ct-xcov: a ct-scan based explainable framework for covid-19 diagnosis. In *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pages 1–8. IEEE, 2023. 3, 4, 2

[33] 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*, pages 1657–1664, 2013. 2

[34] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020. 1

[35] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1

[36] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2

[37] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015. 2, 1

[38] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 7

[39] 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*, pages 611–623. Springer, 2021. 3, 2

[40] Zhongyi Han, Guanglin Zhou, Rundong He, Jindong Wang, Xing Xie, Tailin Wu, Yilong Yin, Salman Khan, Lina Yao, Tongliang Liu, et al. How well does gpt-4v (ision) adapt to distribution shifts? a preliminary investigation. *arXiv preprint arXiv:2312.07424*, 2023. 1, 3

[41] Gregory Holste, Song Wang, Ziyu Jiang, Thomas C Shen, George Shih, Ronald M Summers, Yifan Peng, and Zhangyang Wang. Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In *MICCAI Workshop on Data Augmentation, Labelling, and Imperfections*, pages 22–32. Springer, 2022. 4

[42] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis, 2019. 1

[43] Hossein Jafarinia, Alireza Alipanah, Danial Hamdi, Saeed Razavi, Nahal Mirzaie, and Mohammad Hossein Rohban. Snuffy: Efficient whole slide image classifier. *arXiv preprint arXiv:2408.08258*, 2024. 4

[44] Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, and et al. Drugood: Out-of-distribution (ood) dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. *arXiv preprint arXiv:2201.09637*, 2022. 3, 4

[45] Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers, 2024. 1

[46] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, and et al. Wilds: A benchmark of in-the-wild distribution shifts, 2021. 1

[47] Zhangli Lan, Songbai Cai, Xu He, and Xinpeng Wen. Fixcaps: An improved capsules network for diagnosis of skin cancer. *IEEE Access*, 10:76261–76267, 2022. 4

[48] Gyuseong Lee, Wooseok Jang, Jin Hyeon Kim, Jaewoo Jung, and Seungryong Kim. Domain generalization using large pretrained models with mixture-of-adapters. *arXiv preprint arXiv:2310.11031*, 2023. 3

[49] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, and et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Information Processing Systems*, 33:9459–9474, 2020. 6

[50] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning, 2023. 1

[51] 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*, pages 5542–5550, 2017. 2

[52] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M. Hospedales. Episodic training for domain generalization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1446–1455, 2019. 1

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

[54] Ziyue Li, Kan Ren, Xinyang Jiang, Yifei Shen, Haipeng Zhang, and Dongsheng Li. Simple: Specialized model-sample matching for domain generalization. In *International Conference on Learning Representations*, 2023. 1

[55] Ziyue Li, Kan Ren, Xinyang Jiang, Yifei Shen, Haipeng Zhang, and Dongsheng Li. Simple: Specialized model-sample matching for domain generalization. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[56] Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. Zin: When and how to learn invariance without environment partition? *Advances in Neural Information Processing Systems*, 35: 24529–24542, 2022. 8

[57] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 1

[58] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 3

[59] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *arXiv e-prints*, pages arXiv–2310, 2023. 1

[60] Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Robust place categorization with deep domain generalization. *IEEE Robotics and Automation Letters*, 3:2093–2100, 2018. 1

[61] Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Best sources forward: Domain generalization through source-specific nets. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1353–1357, 2018. 1

[62] Xiaofeng Mao, Yufeng Chen, Xiaojun Jia, Rong Zhang, Hui Xue, and Zhao Li. Context-aware robust fine-tuning. *International Journal of Computer Vision*, 132(5):1685–1700, 2024. 3

[63] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. Chembl: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019. 3

[64] National Institutes of Health et al. Nih clinical center provides one of the largest publicly available chest x-ray datasets to scientific community, 2017. 3

[65] OpenAI. Gpt-4 technical report, 2023. 1, 2, 5

[66] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4, 2023. 1

[67] 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*, pages 1406–1415, 2019. 2

[68] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2

[69] Vihari Piratla, Praneeth Netrapalli, and Sunita Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *International Conference on Machine Learning*, 2020. 1

[70] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12553–12562, 2020. 1

[71] 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*, pages 8748–8763. PMLR, 2021. 2

[72] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Recycling diverse models for out-of-distribution generalization. *ArXiv*, abs/2212.10445, 2022. 1

[73] Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *ArXiv*, abs/2205.09739, 2022. 1

[74] Alexandre Ramé, Kartik Ahuja, Jianyu Zhang, Matthieu Cord, Léon Bottou, and David Lopez-Paz. Model ratatouille: Recycling diverse models for out-of-distribution generalization. In *International Conference on Machine Learning*, pages 28656–28679. PMLR, 2023. 3

[75] Seonguk Seo, Yumin Suh, Dongwan Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. *ArXiv*, abs/1907.04275, 2019. 1

[76] Zhengyang Shen, Lingshen He, Zhouchen Lin, and Jinwen Ma. Pdo-econvs: Partial differential operator based equivariant convolutions. In *International Conference on Machine Learning*, pages 8697–8706. PMLR, 2020. 3

[77] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023. 2

[78] Tianbo Sun, Tong Meng, Yutong Liu, et al. Camelyon 17 challenge: A comparison of traditional machine learning (svm) with the deep learning method. *Wireless Communications and Mobile Computing*, 2022, 2022. 3, 2

[79] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, and et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2

[80] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, and et al. Llama: Open and efficient foundation language models, 2023. 1

[81] 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. 3

[82] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models, 2021. 1

[83] 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, pages 5018–5027, 2017. 2

[84] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023. 1

[85] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2

[86] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning, 2023. 1

[87] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context, 2023. 1

[88] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, and et al. Emergent abilities of large language models, 2022. 1

[89] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, and et al. Chain-of-thought prompting elicits reasoning in large language models, 2023. 1

[90] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, and et al. Larger language models do in-context learning differently, 2023. 1

[91] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, and et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *ArXiv*, abs/2203.05482, 2022. 1

[92] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14378–14387, 2021. 1

[93] Renzhe Xu, Peng Cui, Zheyan Shen, Xingxuan Zhang, and Tong Zhang. Why stable learning works? a theory of covariate shift generalization. *ArXiv*, abs/2111.02355, 2021. 1

[94] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, and et al. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 1

[95] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2

[96] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations. *arXiv preprint arXiv:2306.04618*, 2023. 1

[97] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Scholkopf, and Eric P. Xing. Towards principled disentanglement for domain generalization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8014–8024, 2021. 1

[98] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 2

[99] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, and et al. Llama-adapter: Efficient fine-tuning of language models with zero-init attention, 2023. 1

[100] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, and et al. Opt: Open pre-trained transformer language models, 2022. 1

[101] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5368–5378, 2021. 1

[102] Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, and Peng Cui. Nico++: Towards better benchmarking for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16036–16047, 2023. 2, 3

[103] Xingxuan Zhang, Renzhe Xu, Han Yu, Yancheng Dong, Pengfei Tian, and Peng Cui. Flatness-aware minimization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5189–5202, 2023. 1

[104] Xingxuan Zhang, Renzhe Xu, Han Yu, Hao Zou, and Peng Cui. Gradient norm aware minimization seeks first-order flatness and improves generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20247–20257, 2023. 1

[105] Yiyuan Zhang, Kaixiong Gong, Xiaohan Ding, Kaipeng Zhang, Fangrui Lv, Kurt Keutzer, and Xiangyu Yue. Towards unified and effective domain generalization. *arXiv preprint arXiv:2310.10008*, 2023. 3

[106] YiFan Zhang, Xue Wang, Jian Liang, Zhang Zhang, Liang Wang, and et al. Free lunch for domain adversarial training: Environment label smoothing, 2023. 1

[107] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *arXiv preprint arXiv:2301.13670*, 2023. 5

[108] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. *ArXiv*, abs/2003.06054, 2020. 1

[109] Kaiyang Zhou, Yongxin Yang, Timothy M. Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. *ArXiv*, abs/2007.03304, 2020. 1

[110] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2