Adapting Large Multimodal Models to Distribution Shifts: The Role of In-Context Learning

Guanglin Zhou[1†*], **Zhongyi Han**[2†], **Shiming Chen**[3], **Biwei Huang**[4],
Liming Zhu[5], **Salman Khan**[3,6], **Xin Gao**[2*], **Lina Yao**[5,1,7*]
[1]University of New South Wales, [2]King Abdullah University of Science and Technology,
[3]Mohamed bin Zayed University of AI, [4]University of California, San Diego,
[5]Data61, CSIRO, [6]Australian National University, [7]Macquarie University
jameszhou.ustc@gmail.com

Abstract

Recent studies indicate that large multimodal models (LMMs) are highly robust against natural distribution shifts, often surpassing previous baselines. Despite this, domain-specific adaptation is still necessary, particularly in specialized areas like healthcare. Due to the impracticality of fine-tuning LMMs given their vast parameter space, this work investigates *in-context learning* (ICL) as an effective alternative for enhancing LMMs' adaptability. We find that the success of ICL heavily relies on the choice of demonstration, mirroring challenges seen in large language models but introducing unique complexities for LMMs facing distribution shifts. Our study addresses this by evaluating an unsupervised ICL method, TopKNearestPR, which selects in-context examples through a nearest example search based on feature similarity. We uncover that its effectiveness is limited by the deficiencies of pre-trained vision encoders under distribution shift scenarios, evidenced by their zero-shot capabilities barely outperforming random guesses. To address these challenges, we propose InvariantSelectPR, a novel method leveraging Class-conditioned Contrastive Invariance (CCI) for more robust demonstration selection. Specifically, CCI enhances pre-trained vision encoders by improving their discriminative capabilities across different classes and ensuring invariance to domain-specific variations. This enhancement allows the encoders to effectively identify and retrieve the most informative examples, which are then used to guide LMMs in adapting to new query samples under varying distributions. Our experiments show that InvariantSelectPR substantially improves the adaptability of LMMs, achieving significant performance gains on benchmark datasets, with a 34.2%↑ accuracy increase in 7-shot on Camelyon17 and 16.9%↑ increase in 7-shot on HAM10000 compared to the baseline zero-shot performance. Our code will be publicly available at: https://github.com/jameszhou-gl/icl-distribution-shift.

1 Introduction

Machine learning models are essential in areas such as climate modeling, biomedicine, and autonomous driving, where they need to reliably manage deviations from their training data known as distribution shifts [Park et al., 2021, Zhou et al., 2022, 2023b]. Traditional methods like domain adaptation and domain generalization have been somewhat effective but still fall short in addressing these shifts, as confirmed by several empirical studies [Gulrajani and Lopez-Paz, 2020, Wiles et al., 2022]. However, the emergence of foundation models, characterized by their extensive and diverse
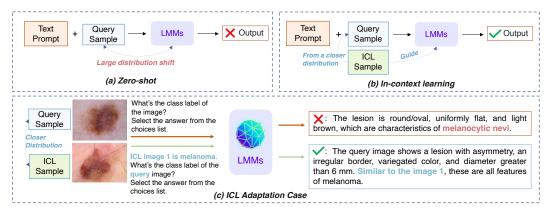

Figure 1: Comparative illustration of (a) zero-shot transfer, which relies on LMMs' pre-trained knowledge to respond to queries, potentially leading to a large distribution gap, and (b) in-context learning (ICL), which introduces an example from a closer distribution to bridge this gap. This work investigates different retrieval methods for selecting effective ICL examples. (c) The efficacy of one-shot ICL is showcased in guiding LMMs with shifted distributions.

pretraining, offers new possibilities for enhancing adaptability to these challenges [Bommasani et al., 2021, Radford et al., 2021, Shu et al., 2023]. Specifically, large multimodal models (LMMs) [Yang et al., 2023b] such as GPT-4V [OpenAI, 2023], and Gemini [Team et al., 2023] have shown superior adaptability. Their zero-shot[2] capabilities have been found to frequently outperform the performance of traditional fine-tuned models in natural datasets [Han et al., 2024].

Despite recent advances, domain-specific adaptation remains a significant challenge, especially in healthcare and scientific research [Han et al., 2024]. While LMMs like Google DeepMind's Med-Gemini offer fine-tuned versions for medical tasks [Saab et al., 2024], their block-box nature and massive parameter sets make traditional fine-tuning impractical for researchers without extensive computational resources. This highlights the urgent need for more feasible adaptation techniques. *In-context learning* (ICL), which allows models to adapt during inference without parameter adjustments, emerges as a promising alternative [Brown et al., 2020, Liu et al., 2021, Min et al., 2022, Dong et al., 2022]. While the effectiveness of ICL is recognized within large language models (LLMs), its application for improving adaptability in LMMs under distribution shifts is less explored.

As depicted in Figure 1, we hypothesize that equipping LMMs with context examples that include task-specific information and details about the query sample can substantially enhance their performance. Our research starts with a thorough evaluation of ICL's capacity to tailor LMMs (§2) to specific domains, particularly healthcare research, where there is a clear necessity for domain-specific fine-tuning [Han et al., 2024]. We discover that the success of ICL heavily depends on the choice of demonstrations, supporting findings from previous LLMs studies [Liu et al., 2021, Min et al., 2022, Dong et al., 2022]. Although the significant impact of demonstration selection on ICL performance is not entirely unexpected, the challenge of selecting demonstrations for LMMs under distribution shifts remains unexplored. To address this, we re-examine the unsupervised retrieval of in-context examples (§3.1), TopKNearestPR, traditionally used in LLMs. This method, discussed in [Liu et al., 2021, Zhang et al., 2024], uses feature similarity to pinpoint contextually relevant ICL examples.

However, the TopKNearestPR approach faces considerable challenges when applied to LMMs during distribution shifts. Notably, using pretrained vision encoders like CLIP-ViT[3], zero-shot performance often remains at levels comparable to random guessing in specialized domains, as shown in a recent study [Han et al., 2024]. This poor performance reveals a critical limitation in these encoders: they struggle to recognize and adapt to the subtle variations in new distributions, which compromises the reliability of visual feature similarities for selecting effective demonstrations. To tackle these challenges, we propose InvariantSelectPR, a novel method designed specifically for scenarios involving distribution shifts (§3.2). This approach employs Class-conditioned Contrastive

---

[2]We adopt the zero-shot setting in the CLIP study [Radford et al., 2021], which enables zero-shot transfer via paired image-text features without extra linear probes. It differs from the classical zero-shot of generalizing to unseen categories [Chen et al., 2021, 2022, 2023].

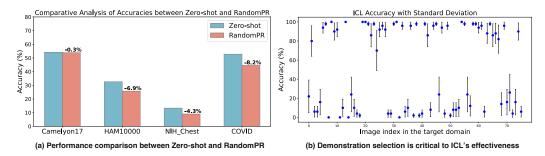[3]https://huggingface.co/openai/clip-vit-base-patch16

Figure 2: Motivation illustration: (a) Performance comparison between Zero-shot and RandomPR, illustrating the limitations of random in-context example selection across four datasets, where one-shot RandomPR often underperforms compared to zero-shot. (b) Analysis of 77 query samples from the target domain, `hospital_3` in Camelyon17, using 50 distinct one-shot examples to examine performance variability. Mean values are marked in blue, and variance is represented by black lines, highlighting the significant impact of example selection on model accuracy. If appropriate in-context samples are chosen, there is a potential for gains up to 40.25%.

Invariance (CCI) to choose demonstrations based on domain-invariant features, which are inherently robust to distributional changes [Zhou et al., 2024]. This retriever method is distinctively crafted for distribution shifts, ensuring the resilience of selected in-context examples in varying conditions. Our empirical results demonstrate that InvariantSelectPR significantly improves the adaptability of LMMs, achieving notable accuracy improvements, *i.e.*, a 34.2% accuracy improvement in 7-shot on Camelyon17 and a 16.9% accuracy increase in 7-shot on HAM10000 over the zero-shot baseline.

Our contributions and the key findings are summarized as follows: (1) To the best of our knowledge, this work takes the first step towards deeply understanding the role of in-context learning as an effective strategy for enhancing the adaptability of LMMs under distribution shifts (§2). (2) We introduce InvariantSelectPR, a novel in-context retrieval framework specifically developed to tackle distribution shifts (§3). (3) Through extensive experiments on four benchmark datasets (§4), our InvariantSelectPR method shows substantial enhancements over zero-shot generalization capabilities.

2   Motivation: ICL Demonstrations under Distribution Shifts

In this section, we evaluate ICL's capability to enhance LMM adaptability. Starting with RandomPR, we randomly select in-context examples from source domain data without relevance to the target task. Our evaluation uses the Gemini model, noted for its zero-shot capabilities, across four medical datasets typically needing domain-specific fine-tuning [Han et al., 2024]. We compare one-shot RandomPR with the zero-shot baseline for a preliminary investigation. According to Figure 2(a), while RandomPR presents a slight decrease of 0.3% on the Camelyon17 dataset, it leads to a substantial performance decline of 4.2%, 3.7%, and even 8.2% on the HAM10000, NIH_Chest, and COVID datasets respectively. Despite its conceptual simplicity, our empirical results suggest that random in-context example selection often fails to fulfill the essential requirement for effective model adaptation—providing informative and contextually appropriate demonstrations.

To unravel the variable efficacy of RandomPR, we conduct an experiment using the Camelyon17 dataset, focusing on 77 query samples from the target domain `hospital_3`. We test the influence of introducing 50 distinct examples from the source domains on the predictions for each query sample. The results in Figure 2(b), display both the mean and standard deviation, indicating significant variability in performance based on the in-context examples used. Notably, while zero-shot accuracy is 54.55% ($44/77$), our analysis reveals that up to 73 query samples could be accurately classified with the apt in-context samples, potentially boosting accuracy by 40.25%. Furthermore, the variability observed—such as a mean accuracy of 70% and a 21% variance for the 25th query—highlights the varying effects of different ICL examples. These findings highlight the inconsistencies in RandomPR's performance and underscore the need for advanced methodologies in ICL example selection.
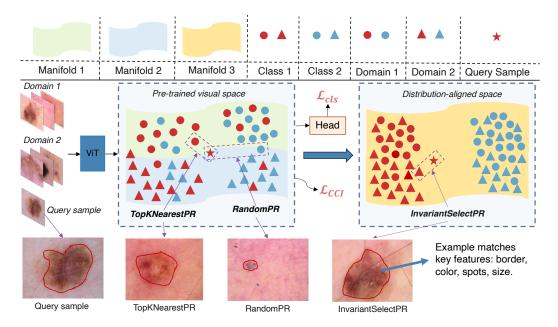
3

Figure 3: Overview of three retrieval methods: RandomPR, TopKNearestPR, and InvariantSelectPR. RandomPR selects examples without specific criteria, often overlooking informative ones. TopKNearestPR uses feature similarities for selection, yet struggles with domain-specific tasks where pre-trained encoder features lack sufficient detail. In contrast, InvariantSelectPR uses a class-conditioned contrastive invariance (CCI) framework to enhance vision encoders, effectively identifying the most representative samples by focusing on key invariant features.

3 Methodology

Upon identifying the limitations of RandomPR, we developed two advanced methods for more effective ICL example selection: TopKNearestPR and InvariantSelectPR, illustrated in Figure 3. These methods aim to enhance the adaptability of LMMs to distribution shifts through the strategic selection of demonstrative examples. We detail these selection methods below.

3.1 TopKNearestPR: Enhancing Context Relevance

TopKNearestPR adopts an unsupervised strategy to identify in-context examples by measuring the similarity between the feature vectors of a target query image $\boldsymbol{x}_q$ and those across M source domains. The dataset $\mathcal{S}$ includes domains $\mathcal{S}^i = \{(\boldsymbol{x}_j^i, \boldsymbol{y}_j^i)\}_{j=1}^{n_i}$, where $\boldsymbol{x}$ represents feature vectors and $\boldsymbol{y}$ is class labels. The cosine similarity between the feature vectors from the query image $\boldsymbol{x}_q$ and any image $\boldsymbol{x}_j^i$ from the dataset, calculated using a pre-trained vision encoder like CLIP-ViT, is given by:

$$\mathrm{sim}(\boldsymbol{x}_q, \boldsymbol{x}_j^i) = \frac{\boldsymbol{z}(\boldsymbol{x}_q) \cdot \boldsymbol{z}(\boldsymbol{x}_j^i)}{\|\boldsymbol{z}(\boldsymbol{x}_q)\| \|\boldsymbol{z}(\boldsymbol{x}_j^i)\|} \tag{1}$$

Here, $\boldsymbol{z}(\boldsymbol{x})$ refers to the feature vector extracted by the encoder. The top K images that exhibit the highest similarity to the query are selected using:

$$\mathrm{top}_K \left(\{\mathrm{sim}(\boldsymbol{x}_q, \boldsymbol{x}_j^i) : i = 1, \ldots, M; j = 1, \ldots, n_i\} \right) \tag{2}$$

where top_K denotes the operation of selecting the indices of the K largest values from the set. The selected images serve as the in-context examples for the LMMs, aiming to enhance their understanding and performance on analogous tasks without further training.

3.2 InvariantSelectPR: Tailored for Distribution Shift Adaptation

TopKNearestPR focuses on relevance by utilizing feature similarities, but its effectiveness can be constrained by the granularity of features from conventional encoders. Pretrained vision encoders,

4

such as CLIP-ViT, while robust in general scenarios, often struggle to differentiate effectively in domain-specific tasks. This limitation manifests as zero-shot performances that are only marginally better than random guesses, leading to the selection of suboptimal in-context examples when relying solely on pre-trained models. Thus, we propose InvariantSelectPR, a new method designed to enhance robustness across distribution shifts.

Facilitating Class-conditioned Contrastive Invariance.   InvariantSelectPR is centered around the Class-conditioned Contrastive Invariance (CCI) mechanism, which aims to improve the model's ability to distinguish between classes while maintaining stability across domain-specific variations [Zhou et al., 2024]. This is achieved by promoting similarity among instances of the same class from different domains and highlighting differences between classes. Using the class token embedding [CLS], x_N, from the final vision transformer (ViT) layer, the CCI loss is defined as:

$$\mathcal{L}_{\text{CCI}} = -\mathbb{E}\left[\log \frac{\exp(z_N \cdot z_{N'}/\tau)}{\sum_{k \neq N} \exp(z_N \cdot z_k/\tau)}\right] \tag{3}$$

Here, $z_{N'}$ is a positive sample of z_N from the same class but possibly a different domain, and z_k signifies a negative sample of z_N from a different class. τ denotes the temperature parameter in contrastive learning [Chen et al., 2020, Zhou et al., 2023a]. This formulation ensures that the learned representations are both discriminative and invariant, crucial for adapting to new distributions.

This approach combines this CCI loss $\mathcal{L}_{\text{CCI}}$ with a classification loss $\mathcal{L}_{\text{cls}}$ to enhance the vision encoder's ability to manage distribution shifts effectively. The classification loss uses cross-entropy to align the final class token embedding $\mathbf{x}_N$ with the ground-truth label y, bolstering the model's discriminative power:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{C} y_i \log(\text{Head}(x_N)_i) \tag{4}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{CCI}} \tag{5}$$

where C represents the total number of classes in the dataset, and $\text{Head}(\cdot)$ is a neural classification head that maps the class token x_N to a predicted class probability distribution. λ is a tuning hyper-parameter to control the weight of the CCI loss.

In-Context Selection Through Enhanced Invariance.   After fine-tuning the vision encoder with the combined losses in Eq. (5), we leverage refined features to assess the similarity between the target samples and in-context examples. By ensuring these similarities reflect both visual resemblance and domain invariance, the k-shot examples with the highest similarity scores are then selected.

4   Experiments

Datasets Overview.   We use four benchmark datasets to explore distribution shifts, particularly emphasizing domain-specific fine-tuning [Han et al., 2024]. Camelyon17 [Bandi et al., 2018] features 450,000 patches from breast cancer images across five hospitals. HAM10000 [Tschandl et al., 2018] offers dermatoscopic images critical for skin cancer detection. The NIH_Chest dataset [Wang et al., 2017] includes over 112,000 X-ray images annotated for thoracic diseases. The COVID dataset [Han et al., 2021] provides diverse pneumonia detection data, including COVID-19 cases, from various hospitals. We analyze a practical subset, `random_1`, with 450 samples[4].

Implementation Details.   We compare three retrieval methods—RandomPR, TopKNearestPR and InvariantSelectPR—against the baseline zero-shot capability.   We employ `vit_large_patch14_224_clip_laion2b` configuration from the timm library, exploring variations in backbone configurations further in §4.3.2. The Gemini model is employed as the primary LMM due to its superior zero-shot performance across varied datasets [Han et al., 2024] and its stable log-linear improvement in performance with an increasing number of ICL examples, as observed in a concurrent study [Jiang et al., 2024]. Our main results (§4.1) focus on one-shot performance, with additional insights on the impact of different numbers of shots in §4.3.3. We also include other leading LMMs, such as GPT-4V and Claude [Anthropic, 2023], in our extended analysis in §4.3.4.

---

[4]Available at https://github.com/jameszhou-gl/gpt-4v-distribution-shift

Table 1: Performance comparison of three retrieval methods against the zero-shot approach, illustrating accuracy improvements or decreases on Camelyon17 and COVID datasets. Mean and standard deviation values are calculated over three independent runs, with the best results highlighted in bold.

Method	Camelyon17						COVID		
	Hosp0	Hosp1	Hosp2	Hosp3	Hosp4	Acc	Sou	Tar	Acc
Zero-shot	52.00	51.93	56.44	54.55	56.67	54.17±0.5	62.19	44.19	52.75±1.3
RandomPR	50.50	53.03	54.59	53.28	58.55	53.87±1.1	38.66	49.86	44.52±0.9
TopKNearestPR	62.24	58.65	58.62	59.65	60.15	59.91±1.8	41.59	60.88	51.70±2.7
InvariantSelectPR	60.12	63.96	62.68	63.39	64.36	**62.77±1.1**	39.94	67.05	**54.15±1.0**

Table 2: Performance comparison of three retrieval methods against the zero-shot approach, illustrating accuracy improvements or decreases on HAM10000 and NIH_Chest datasets. Mean and standard deviation values are calculated over three independent runs, with the best results highlighted in bold.

Method	HAM10000					NIH_Chest		
	RD	VMod	VMol	VDis	Acc	PA	AP	Acc
Zero-shot	28.54	37.39	26.42	42.22	32.62±1.2	12.41	14.26	13.31±0.7
RandomPR	21.48	32.81	27.88	12.64	25.71±1.9	7.91	10.12	8.98±2.1
TopKNearestPR	23.03	33.39	31.73	28.03	28.72±1.0	10.34	10.43	10.39±1.0
InvariantSelectPR	38.20	49.43	38.96	25.19	**40.91±1.0**	13.22	13.63	**13.42±0.7**

Training Protocol. We train InvariantSelectPR for 100 epoches using the AdamW optimizer [Loshchilov and Hutter, 2017], with a learning rate of $1e$-5 and a weight decay of 0.01. For clarity, both the temperature parameter τ and the loss weight λ are fixed at 1.0. Each dataset is specifically fine-tuned to optimize the vision encoder for its respective domains. Experiments are conducted on a Linux server equipped with an Intel Xeon CPU, NVIDIA A5000 and V100 GPUs.

4.1 Main Results

In Tables 1 and 2, our analysis of four benchmark datasets provides a detailed examination of how different ICL methods perform under distribution shifts. The zero-shot approach highlights the inherent ability of LMMs to adapt to new domains without retraining. However, the effectiveness of this adaptability varies significantly with different ICL strategies. The RandomPR strategy, which employs a stochastic method for selecting in-context examples, yields inconsistent results. For instance, on the Camelyon17 dataset, it leads to a slight decrease of 0.3% in accuracy, but it largely underperforms on the COVID, HAM10000, and NIH_Chest datasets, with accuracy decreases of 8.2%, 4.2%, and 3.7%, respectively. This highlights the unpredictable performance of RandomPR across different conditions. Conversely, TopKNearestPR uses a pre-trained vision encoder to identify feature similarities for example selection, leading to a 5.74% improvement on the Camelyon17, which demonstrates the benefits of a more targeted approach in example selection. Despite this success, the method sees declines of 3.9% and 2.9% on the HAM10000 and NIH_Chest datasets, respectively, indicating a lack of consistent performance across all test scenarios. The most effective strategy, InvariantSelectPR, consistently outperforms other methods, significantly exceeding the zero-shot baseline across all datasets, especially achieving remarkable gains of 8.3% on HAM10000 and 8.6% on Camelyon17. These results underscore the importance of advanced in-context example selection techniques in adapting LMMs to distribution shifts. Despite notable gains, the improvements with InvariantSelectPR on NIH_Chest and COVID are modest. In Figure 4, the incremental improvements by InvariantSelectPR align with those from fine-tuned encoders, which generally surpass the fine-tuning approach by 1% to 6%. This suggests that when fine-tuning itself is minimally effective, ICL strategies yield limited enhancements. Future research thus focuses on more sophisticated methods to enhance invariance, beyond fundamental domain-invariance in this work.

4.2 Ablation Study

To assess the impact of enhanced invariance on model adaptability, we conduct an ablation study focusing on the Gemini model's one-shot performance. This study compares three configurations: a baseline using TopKNearestPR, the baseline only with $\mathcal{L}_{cls}$, and the full InvariantSelectPR that incorporates both $\mathcal{L}_{cls}$ and $\mathcal{L}_{CCI}$. Table 3 displays incremental performance gains across datasets with the successive additions of $\mathcal{L}_{cls}$ and $\mathcal{L}_{CCI}$. The addition of $\mathcal{L}_{cls}$ alone leads to a modest increase in performance by 1.77%. However, when incorporated with $\mathcal{L}_{CCI}$, there is a more substantial performance boost of 4.01%, confirming the effectiveness of CCI loss in improving the models' adaptability.

Table 3: Ablation study on loss terms. The baseline is TopKNearestPR.

Configurations	Camelyon17	COVID	HAM10000	NIH_Chest	**Average**
baseline	61.96	54.22	29.84	10.49	39.13
baseline+$\mathcal{L}_{cls}$ (w/o CCI)	61.59	52.00	38.93	11.11	40.90
baseline+$\mathcal{L}_{cls}$+$\mathcal{L}_{CCI}$ (full)	**63.90**	**54.44**	**41.56**	**12.67**	**43.14**

4.3 In-depth Analysis

4.3.1 ICL vs. Traditional Supervised Finetuning (SFT)

Recent advances in LMMs demonstrate their impressive zero-shot generalization, often outperforming fine-tuned models in natural distribution shifts [Han et al., 2024]. This raises questions about whether LMMs with in-context learning, can exceed fine-tuned model performance in scientific datasets, traditionally reliant on domain-specific fine-tuning. We assess our ICL strategy against traditional SFT to explore this. For SFT, we focus on maintaining domain invariance and targeting the class prediction objective, similar to Eq. (5). We fine-tune a CLIP-ViT on source domain data and then apply it to predict outcomes on target examples. This comparison directly measures the effectiveness of ICL versus conventional SFT. For InvariantSelectPR, we choose ICL examples ranging from one to seven and report the best accuracy. Figure 4 displays the comparative performance across four datasets. SFT demonstrates a substantial improvement over zero-shot capabilities with accuracy improvements of 32.4%, and 12.3% on Camelyon17 and HAM10000, but underperforms 1.5% and 5.3% on NIH_Chest and COVID. In contrast, our proposed InvariantSelectPR with few-shot examples, consistently exceeds SFT, with gains of 1.4%, 4.4%, 1.6%, and 6.4% in the same datasets.
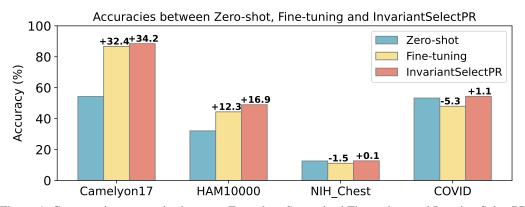

Figure 4: Comparative accuracies between Zero-shot, Supervised Finetuning, and InvariantSelectPR methods across various datasets, illustrating the superior performance of InvariantSelectPR over both zero-shot and supervised fine-tuning.

4.3.2 Backbone Evaluation

We evaluate the impact of different vision encoder backbones on the effectiveness of our InvariantSelectPR method compared to TopKNearestPR. This includes ViT models pretrained on ImageNet-21K

and trained with the self-supervised DINO method on ImageNet-1K. Table 4 shows that InvariantSelectPR consistently outperforms TopKNearestPR across datasets, with an average accuracy of 42.9% versus 39.7%. This underscores the limitations of relying solely on pretrained visual similarity for selecting meaningful in-context examples. InvariantSelectPR also demonstrates more consistent performance, with less deviation from mean accuracy (under 1%) compared to TopKNearestPR (nearly 2%). An important observation is the enhanced performance of backbones utilizing self-supervised or contrastive learning methods, supporting the effectiveness of self-supervised learning in capturing generalizable features that contribute to more robust ICL performance, as suggested in studies [Chen et al., 2020, Radford et al., 2021, Wang et al., 2023a].

Table 4: Performance comparison of ICL methods with different vision encoder backbones.

Methods	Backbones	Camelyon17	COVID	HAM10000	NIH_Chest	**Average**
TopKNearestPR	vit-l/14-clip	61.96	54.22	29.84	10.49	39.13
	vit-l/16-in21k	60.45	49.78	30.80	10.24	37.82
	vit-b/16-dino	63.72	59.68	32.74	10.89	41.76
InvariantSelectPR	vit-l/14-clip	63.90	54.44	41.56	12.67	43.14
	vit-l/16-in21k	61.76	52.78	38.15	12.89	41.40
	vit-b/16-dino	64.48	53.72	44.55	11.36	43.53

4.3.3 ICL Examples with Various Shots

To assess the impact of the number of ICL examples, we perform an empirical study using the Camelyon17 and HAM10000 datasets, varying the number of shots from 1 to 7 for each dataset in Figure 5. This analysis reveals that increasing the number of shots leads to a decrease in the performance of the RandomPR method, implying that additional examples might introduce unhelpful information. In contrast, the TopKNearestPR method typically improves with more shots but shows a decline in performance when moving from 3-shot to 5-shot on the HAM10000 dataset, suggesting potential issues with example selection or redundancy. On the other hand, our InvariantSelectPR method consistently improves performance as the number of shots increases, demonstrating its effectiveness in utilizing information from source domains. Notably, this method achieves a performance boost of approximately 24.6% when the shot count increases from 1 to 7 on Camelyon17.
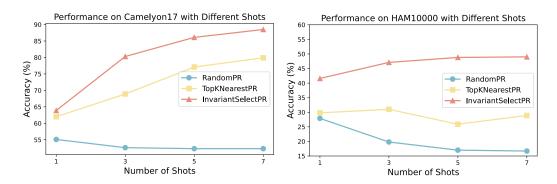

Figure 5: Performance comparison with varying numbers of ICL examples (shots) on Camelyon17 and HAM10000 datasets.

4.3.4 Evaluation Across Different LMMs

The open-source LMMs like IDEFICS [Laurençon et al., 2024] and OpenFlamingo [Awadalla et al., 2023] primarily focus on text and ignore the input signal of images [Bertini Baldassini et al., 2024]. Furthermore, these LMMs lack instruction-following ability to choose the response from the answer list. Thus, we use three proprietary LMMs in this comparative analysis: Gemini Pro, GPT-4V, and Claude 3 Opus [Anthropic, 2024]. Due to the high computational demands and associated costs of GPT-4V and Claude 3 Opus, we limit our testing to a single dataset, HAM10000, and perform a one-shot evaluation. Figure 6 demonstrates that InvariantSelectPR consistently outperforms other

methods across all three LMMs. This method not only exceeds baseline zero-shot performance but also significantly enhances adaptability. Both GPT-4V and Claude 3 Opus exhibit substantial improvements using all ICL methods over their zero-shot capabilities, suggesting that ICL can effectively boost the adaptability of LMMs. This analysis highlights the capacity of InvariantSelectPR to leverage domain-invariant features to enhance LMMs performance under variable conditions.
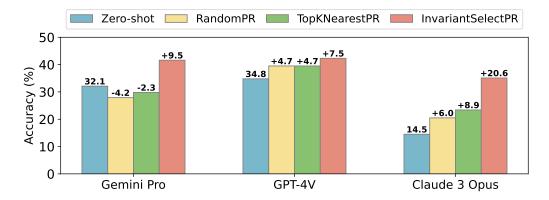

Figure 6: Performance comparison of Zero-shot, RandomPR, TopKNearestPR, and InvariantSelectPR on the HAM10000 dataset across three LMMs, demonstrating the impact of one-shot demonstrations.

4.3.5   Distance Metric Evaluation

We examine the effects of employing various distance metrics, including Cosine, Euclidean, and Manhattan, within TopKNearestPR and InvariantSelectPR methods, as illustrated in Table 5. Our findings indicate that the InvariantSelectPR method consistently achieves higher performance than TopKNearestPR across all metrics tested on both the Camelyon17 and HAM10000 datasets.

Table 5: Performance comparison using different distance metrics.

Method	Camelyon17				HAM10000			
	Cosine	Euclidean	Manhattan	**Avg**	Cosine	Euclidean	Manhattan	**Avg**
TopKNearestPR	61.96	60.77	59.18	60.64	29.84	29.46	30.80	30.03
InvariantSelectPR	63.90	61.09	62.70	62.56	41.56	37.22	38.93	39.24

4.3.6   Computational Efficiency

Table 6 illustrates the trade-off between computational cost and accuracy improvement. While InvariantSelectPR incurs a slightly higher inference time and GPU usage than the zero-shot baseline but offers an 8.60% accuracy improvement. The increased cost is due to the model loading and similarity calculation. InvariantSelectPR' lower inference time compared to TopKNearestPR is because it loads the vision encoder once per environment instead of for each target sample. Future work will focus on optimizing these steps to reduce inference time while maintaining accuracy gains.

Table 6: Performance comparison of different one-shot ICL methods on the Camelyon17 dataset, in terms of inference time, GPU usage, and accuracy improvement over the zero-shot baseline.

Method	Inference Time (s/query)	GPU Usage (GB)	Accuracy Improvement
Zero-shot	5.23	-	-
RandomPR	5.35	-	-0.30%
TopKNearestPR	15.52	2.41	+5.74%
InvariantSelectPR	11.79	3.55	+8.60%

5   Conclusion and Discussion

Concusion.   We investigated the efficacy of in-context learning (ICL) to improve the adaptability of LMMs to distribution shifts through our novel ICL approach, InvariantSelectPR. This method not only outperforms standard zero-shot capabilities but also exceeds other methods like RandomPR and TopKNearestPR in handling domain-specific shifts. Evaluations across four datasets confirmed that InvariantSelectPR enhances LMM adaptability by optimally selecting demonstrative examples. Our study offers insights for future work on distribution shifts in foundation models.

Limitations.   Our study is constrained by several factors. Firstly, we confined our analysis to a small selection of benchmark datasets and relied exclusively on commercial and proprietary models, such as Gemini Pro, GPT-4V, and Claude 3 Opus. The limited availability of comprehensive documentation for these models constrains our understanding of their pre-training data, architecture, and inherent biases. This is critical as some broadly used open-source LMMs can not effectively understand multiple images like Flamingo [Alayrac et al., 2022] and simultaneously follow instructions like LLaVA [Liu et al., 2023], necessitating the use of commercial models. Additionally, the substantial financial and computational resources required to access these proprietary models may restrict further validation and analysis. Secondly, our empirical tests involved just 450 samples, which, despite prior research suggesting stability ranging from 180 to 1800 cases [Han et al., 2024], might not reveal scalability issues or subtle biases in larger datasets. Thirdly, the prevalence of numerous domains in healthcare [Yang et al., 2023a] and scientific research [Ji et al., 2022] presents potential challenges in scaling our method across multiple domains.

Boarder Impacts.   This work introduces a novel ICL method to enhance the adaptability of LMMs to distribution shifts, particularly in scientific and healthcare domains. While this method promises improved model accuracy and reliability, its misuse could amplify biases or yield unreliable outputs from LMMs, leading to negative consequences in these critical fields.

References

J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Anthropic. Model card and evaluations for claude models. https://www-cdn.anthropic.com/files/4zrzovbb/website/bd2a28d2535bfb0494cc8e2a3bf135d2e7523226.pdf, 2023. Accessed: 2024-03-07.

Anthropic. Claude 3 haiku: our fastest model yet. 2024. Available at: https://www.anthropic.com/news/claude-3-haiku.

A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt. Openflamingo, Mar. 2023. URL https://doi.org/10.5281/zenodo.7733589.

Y. Balaji, S. Sankaranarayanan, and R. Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, 2018.

I. Balazevic, D. Steiner, N. Parthasarathy, R. Arandjelović, and O. Henaff. Towards in-context scene understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. 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.

A. Bar, Y. Gandelsman, T. Darrell, A. Globerson, and A. Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022.

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

F. Bertini Baldassini, M. Shukor, M. Cord, L. Soulier, and B. Piwowarski. What makes multimodal in-context learning work? *arXiv e-prints*, pages arXiv–2404, 2024.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019.

P. Chattopadhyay, Y. Balaji, and J. Hoffman. Learning to balance specificity and invariance for in and out of domain generalization. In *ECCV*, 2020.

S. Chen, W. Wang, B. Xia, Q. Peng, X. You, F. Zheng, and L. Shao. Free: Feature refinement for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 122–131, 2021.

S. Chen, Z. Hong, Y. Liu, G.-S. Xie, B. Sun, H. Li, Q. Peng, K. Lu, and X. You. Transzero: Attribute-guided transformer for zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 330–338, 2022.

S. Chen, W. Hou, Z. Hong, X. Ding, Y. Song, X. You, T. Liu, and K. Zhang. Evolving semantic prototype improves generative zero-shot learning. In *International Conference on Machine Learning*, pages 4611–4622. PMLR, 2023.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. URL https://api.semanticscholar.org/CorpusID:225039882.

Z. Fang, X. Li, X. Li, J. M. Buhmann, C. C. Loy, and M. Liu. Explore in-context learning for 3d point cloud understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

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

Z. Han, R. He, T. Li, B. Wei, J. Wang, and Y. 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.

Z. Han, X.-J. Gui, H. Sun, Y. Yin, and S. Li. Towards accurate and robust domain adaptation under multiple noisy environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (5):6460–6479, 2022a.

Z. Han, H. Sun, and Y. Yin. Learning transferable parameters for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 31:6424–6439, 2022b.

Z. Han, G. Zhou, R. He, J. Wang, X. Xie, T. Wu, Y. Yin, S. Khan, L. Yao, T. Liu, et al. How well does gpt-4v (ision) adapt to distribution shifts? a preliminary investigation. *arXiv preprint arXiv:2312.07424*, 2023.

Z. Han, G. Zhou, R. He, J. Wang, T. Wu, Y. Yin, S. Khan, L. Yao, T. Liu, and K. Zhang. How well does GPT-4v(ision) adapt to distribution shifts? a preliminary investigation. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. URL https://openreview.net/forum?id=J8V4EwZkez.

Y. Ji, L. Zhang, J. Wu, B. Wu, L.-K. Huang, T. Xu, Y. Rong, L. Li, J. Ren, D. Xue, 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.

Y. Jiang, J. Irvin, J. H. Wang, M. A. Chaudhry, J. H. Chen, and A. Y. Ng. Many-shot in-context learning in multimodal foundation models, 2024.

H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. Rush, D. Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.

D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018a.

Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018b.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

Z. Mao and Y. Yu. Tuning llms with contrastive alignment instructions for machine translation in unseen, low-resource languages. *arXiv preprint arXiv:2401.05811*, 2024.

S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

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

C. Park, A. Awadalla, T. Kohno, and S. Patel. Reliable and trustworthy machine learning for health using dataset shift detection. *Advances in Neural Information Processing Systems*, 34:3043–3056, 2021.

X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019.

V. Piratla, P. Netrapalli, and S. Sarawagi. Efficient domain generalization via common-specific low-rank decomposition. In *ICML*, 2020.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Y. Shu, X. Guo, J. Wu, X. Wang, J. Wang, and M. Long. CLIPood: Generalizing CLIP to out-of-distributions. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31716–31731. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/shu23a.html.

B. Sun and K. 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*, pages 443–450. Springer, 2016.

G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017.

R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, 2018.

H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023a.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 2097–2106, 2017.

X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023b.

Z. Wang, Y. Jiang, Y. Lu, P. He, W. Chen, Z. Wang, M. Zhou, et al. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022a.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b.

N. Wies, Y. Levine, and A. Shashua. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36, 2024.

O. Wiles, S. Gowal, F. Stimberg, S.-A. Rebuffi, I. Ktena, K. D. Dvijotham, and A. T. Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Dl4LetuLdyK.

Y. Wolf, N. Wies, Y. Levine, and A. Shashua. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023.

J. Wu, T. Yu, R. Wang, Z. Song, R. Zhang, H. Zhao, C. Lu, S. Li, and R. Henao. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

S. M. Xie, A. Raghunathan, P. Liang, and T. Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

C. Yang, M. B. Westover, and J. Sun. ManyDG: Many-domain generalization for healthcare applications. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=lcSfirnflpW.

Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023b.

Y. Zhang, K. Zhou, and Z. Liu. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36, 2024.

G. Zhou, C. Huang, X. Chen, X. Xu, C. Wang, L. Zhu, and L. Yao. Contrastive counterfactual learning for causality-aware interpretable recommender systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3564–3573, 2023a.

G. Zhou, S. Xie, G. Hao, S. Chen, B. Huang, X. Xu, C. Wang, L. Zhu, L. Yao, and K. Zhang. Emerging synergies in causality and deep generative models: A survey. *arXiv preprint arXiv*, 2301, 2023b.

G. Zhou, Z. Han, S. Chen, B. Huang, L. Zhu, T. Liu, L. Yao, and K. Zhang. Hcvp: Leveraging hierarchical contrastive visual prompt for domain generalization. *arXiv preprint arXiv:2401.09716*, 2024.

K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Appendix

More contents are put in the appendix, including:

A Related Work

Distribution Shifts.   The literature on distribution shifts categorizes mitigation approaches into two primary strategies: domain adaptation and domain generalization. Domain adaptation techniques, well-established for scenarios where the target domain is known during training, recalibrate models according to the target data's statistical properties [Ben-David et al., 2006]. These techniques encompass deep transfer learning, which aligns feature distributions between source and target domains [Sun and Saenko, 2016], unsupervised methods that minimize domain discrepancies [Ganin and Lempitsky, 2015], and the use of benchmarks such as Office-Home [Venkateswara et al., 2017] and DomainNet [Peng et al., 2019]. These benchmarks have propelled advances by introducing more complex and diverse scenarios [Han et al., 2022a,b]. In contrast, domain generalization addresses the more daunting challenge of excelling in completely unseen domains. Strategies here include aligning features across multiple source domains [Li et al., 2018b], separating domain-specific from domain-general features [Piratla et al., 2020, Chattopadhyay et al., 2020], employing meta-learning for optimization across various domains [Li et al., 2018a, Balaji et al., 2018], and using data augmentation to mimic domain variability [Volpi et al., 2018, Carlucci et al., 2019]. Recent studies have observed that LLMs have demonstrated exceptional adaptability when dealing with natural distribution shifts but cannot handle the distribution shifts in specialized areas such as healthcare and scientific research [Radford et al., 2021, Han et al., 2023]. This observation motivates this paper's exploration of ICL and the development of new ICL strategies under distribution shifts.

In-Context Learning.   In-context learning (ICL), particularly defined in GPT-3 [Brown et al., 2020], originated in LLMs for natural language processing (NLP) tasks. ICL is a proven effective paradigm that leverages context augmented with a few examples to enable LLMs to make predictions [Dong et al., 2022, Lu et al., 2021, Wei et al., 2022b, Min et al., 2022, Dong et al., 2022, Wei et al., 2022a, Wolf et al., 2023, Wies et al., 2024, Xie et al., 2021]. The choice of these in-context examples critically impacts performance, as evidenced by studies demonstrating that selecting nearest neighbors based on sentence encoders can significantly enhance the few-shot capabilities of models like GPT-3 [Liu et al., 2021, Wu et al., 2024, Mao and Yu, 2024]. While ICL is established in NLP, it is emerging in visual and multimodal LLMs. The study Flamingo [Alayrac et al., 2022] marks the earliest exploration of visual ICL, with subsequent studies validating the importance of example selection in image painting models [Bar et al., 2022, Wang et al., 2023b, Zhang et al., 2024], visual understanding [Balazevic et al., 2024, Fang et al., 2024], and diffusion models [Wang et al., 2024]. Unlike prior work, this paper uniquely focuses on deeply understanding the role of ICL under distribution shifts, taking a first step in this direction.

B Experimental Details

B.1 Datasets

This section provides detailed information about the datasets used in the experiments, including their statistics, preprocessing steps, and domain splits.

Dataset Statistics.   We present the data statistics in Table 7. Besides, we use the datasets available in https://huggingface.co/datasets/jameszhou-gl/gpt-4v-distribution-shift and evaluate random_1 subset with 450 cases.

Preprocessing Steps.   We do not adopt additional preprocessing steps for the images, adhering instead to the guidelines provided by each large multimodal model. For instance, the Gemini model

Table 7: Detailed statistics of the datasets used in the experiments.

Dataset	Category	Prediction Task	Domain Type	# Domains	# Classes	# Samples	Example Image
Camelyon17	Medical	Tumor detection	Hospital	5	2	450	
HAM10000	Medical	Skin disease classification	Hospital	4	7	450	
NIH_Chest	Medical	Lung disease diagnosis	Hospital	2	15	450	
COVID	Medical	Pneumonia type classification	Hospital	2	3	450	

accepts `PIL.Image.open(img_path)` as input without requiring specific image size, normalization, or augmentation[5].

B.2 Prompts

We utilize the following basic prompt template in all ICL experiments, refering to [Han et al., 2024].

```
image_descriptions = [f"Image {i+1} is {image_class}" for i,
                    (desc, image_class) in enumerate(source_images)]

images_description = ". ".join(image_descriptions)

prompt = f"""Given the images, answer the following question, using the
        specified format.
        {images_description}.
        Question: What is the class of the next image?
        Choices: {', '.join(class_names)}.

        Please respond with the following format for each image:
        ---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.
        """
```

---

[5]https://ai.google.dev/gemini-api/docs/get-started/python

B.3 Implementation Details

Vision encoder architectures. We employ the `vit_large_patch14_224_clip_laion2b` configuration from the timm library[6], which utilizes the Vision Transformer (ViT) architecture [Dosovitskiy et al., 2020]. This model divides 224x224 pixel images into 14x14 patches and processes them using a transformer encoder, leveraging extensive pretraining on the LAION-2B dataset with the CLIP approach to enhance visual and textual understanding.

Data and Code. Our code can be found at the supplementary material and data is publicly available in https://huggingface.co/datasets/jameszhou-gl/gpt-4v-distribution-shift.

C t-SNE Visualizations of Visual Features

In this section, we presnet t-SNE [Van der Maaten and Hinton, 2008] visualizations of the visual features to illustrate how class-conditioned contrastive invariance (CCI) contributes to domain invariance and discriminative capabilities. The visualizations are based on the original vision encoder and our fine-tuned vision encoder on three datasets, as shown in Figure 7. The visual features are extracted using both the pretrained and fine-tuned ViT models. The t-SNE plots are created to highlight the clustering behavior of the features from the target domain. By comparing the plots, we can visually assess the impact of the fine-tuning with CCI on the separation of different classes and the compactness of feature clusters.
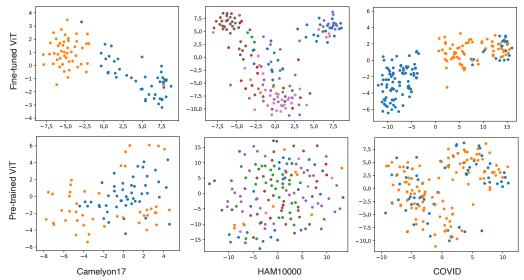

Figure 7: t-SNE visualizations of visual features from the target domain for three datasets. The lower row shows features extracted using the pretrained ViT model, while the upper row shows features extracted using the our fine-tuned ViT model.

These t-SNE visualizations, particularly for the Camelyon17 dataset, clearly demonstrate that fine-tuning with class-conditioned contrastive invariance (CCI) significantly enhances the model's ability to generalize across unseen domains. The fine-tuning process improves discriminative power by better aligning feature representations with class labels. This visualization underscores the critical importance of incorporating CCI to refine the vision encoder. By enhancing the alignment of feature representations with their corresponding classes, CCI contributes to a more robust and domain-invariant model. Furthermore, this refined vision encoder facilitates the selection of in-context learning (ICL) examples, enabling large multimodal models to adapt more effectively.

---

[6] https://github.com/huggingface/pytorch-image-models/blob/main/timm/models/vision_transformer.py