

Delving into Out-of-Distribution Detection with Medical Vision-Language Models

Lie Ju^{1,2,3,4}, Sijin Zhou^{1,4}, Yukun Zhou^{2,3}, Huimin Lu⁵,
Zhuoting Zhu⁶, Pearse A. Keane^{2,3}, and Zongyuan Ge^{1,4}✉

¹ Monash University, Australia

² Moorfields Eye Hospital, United Kingdom

³ University College London, United Kingdom

⁴ Airdoc Technology Inc, China

⁵ Southeast University, China

⁶ CERA, Melbourne University, Australia

{Lie.Ju1, Zongyuan.Ge}@monash.edu

Abstract. Recent advances in medical vision-language models (VLMs) demonstrate impressive performance in image classification tasks, driven by their strong zero-shot generalization capabilities. However, given the high variability and complexity inherent in medical imaging data, the ability of these models to detect out-of-distribution (OOD) data in this domain remains underexplored. In this work, we conduct the first systematic investigation into the OOD detection potential of medical VLMs. We evaluate state-of-the-art VLM-based OOD detection methods across a diverse set of medical VLMs, including both general and domain-specific purposes. To accurately reflect real-world challenges, we introduce a cross-modality evaluation pipeline for benchmarking full-spectrum OOD detection, rigorously assessing model robustness against both semantic shifts and covariate shifts. Furthermore, we propose a novel hierarchical prompt-based method that significantly enhances OOD detection performance. Extensive experiments are conducted to validate the effectiveness of our approach. The codes are available at <https://github.com/PyJulie/Medical-VLMs-OOD-Detection>.

Keywords: Vision Language Models · Out-of-Distribution Detection.

1 Introduction

Recent advances in vision-language models (VLMs), exemplified by CLIP [21], have significantly advanced image recognition through remarkable generalization capabilities, particularly in zero-shot transfer learning. This success has catalyzed a growing interest in the development of medical VLMs, ranging from general-purpose architectures [30,12,9] to domain-specialized experts (e.g., ophthalmology [26,24,22]). Although these models demonstrate high accuracy on in-distribution (ID) samples with textual descriptions, the ability of medical VLMs to distinguish out-of-distribution (OOD) samples remains unclear, especially considering the potential for overconfident predictions on these OOD

samples. In this context, failing to address or assess OOD samples inappropriately can lead to severe outcomes such as misdiagnosis, which could endanger individuals in deployed medical systems [4].

Although many OOD detection methods for image classification have achieved remarkable progress in recent years [10,11,3], conventional vision encoder-only models typically encoded the categories into one-hot vector, leaving the semantic information encapsulated in texts largely unexploited. To utilize the natural advantages of VLMs and address the relevant challenges, some VLM-based methods have been developed for natural images OOD detection, primarily operating under zero-shot [15,18] or few-shot paradigms [16,17]. Zero-shot OOD detection does not require additional training with in-distribution data and typically relies on post-processing techniques. In contrast, few-shot OOD detection involves learning from ID data during both the training and inference phases. To the best of our knowledge, no research has yet explored generalized OOD detection based on medical VLMs, and few studies on medical VLMs have included an analysis of OOD detection capabilities in their experiments. Importantly, it remains unvalidated whether these methods designed for natural image OOD detection are applicable to medical images.

Another major challenge in medical OOD detection is understanding how different types of shifts lead to the exclusion of OOD samples from ID data [4,20]. Previous OOD detection benchmarks primarily focus on identifying outliers with **semantic shifts**, where the semantic labels of OOD samples do not overlap with those of ID samples. For example, predicting natural images using an ophthalmology-specific VLM. Recent OOD detection research has shifted its focus to a more challenging and realistic problem setting: **covariate shifts**. Unlike semantic shifts, covariate shifts do not alter the target categories, meaning OOD samples share the same semantic labels with ID samples but however differ in imaging attributes such as *imaging modalities*, *imaging quality*, or *population distributions* [4]. A typical example is diagnosing lung opacity using a model trained on X-ray images but CT images are incorrectly input. We summarized two types of OOD shifts in Fig. 1, which also reveals that advanced OOD detection techniques often struggle in such scenarios. The evaluation of OOD detection performance across both semantic shifts and covariate shifts is defined as **full-spectrum OOD detection** [28,4].

In this work, we first evaluate state-of-the-art VLM-based OOD detection methods on a set of medical VLMs, including both general-purpose and domain-specific models. Subsequently, we propose a hierarchical prompt-based method that works on both zero-shot inference phase and retraining phase with few-shot fine-tuning. We also define three dataset evaluation pipelines to establish a novel benchmark which simulates challenging conditions for real-world applications.

Based on the above perspectives, the main contributions of this paper are:

1. We present a systematic evaluation of generalized OOD detection capabilities in medical VLMs. By integrating state-of-the-art OOD detection methods and assessing both general-purpose and domain-specific medical VLMs, we

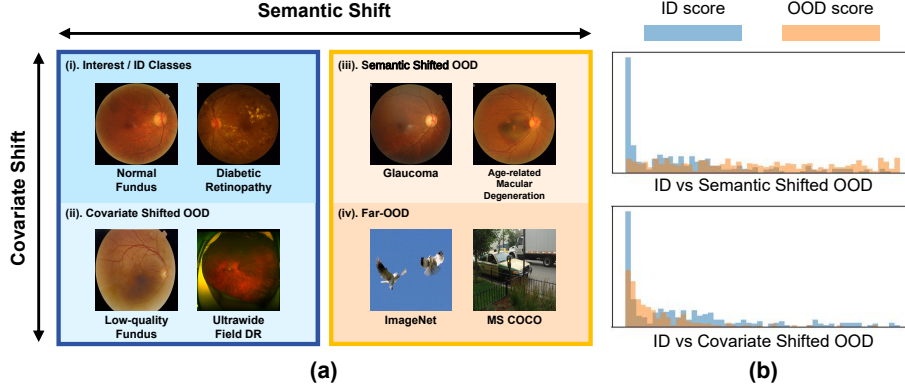


Fig. 1. (a) Problem illustration. (i) ID classes with textual descriptions seen by CLIP-like models are defined as ID classes (e.g., diabetic retinopathy); (ii) Covariate shifted OOD data sharing semantic relevance with ID classes but exhibiting covariate shifts, such as low image quality or differences in imaging devices (e.g., ultrawide-field fundus imaging). (iii) & (iv) OOD with irrelevant concerns of semantics. (b) A simple baseline experiment demonstrates that advanced OOD detection techniques (e.g., MCM [15]) tend to fail on covariate-shifted OOD scenarios.

establish a novel benchmark that rigorously addresses the unique challenges of medical imaging, including semantic and covariate shifts.

2. To bridge the gap between synthetic evaluations and real-world clinical scenarios, we design three benchmark datasets that simulate full-spectrum OOD detection challenges. These datasets encompass diverse imaging modalities, anatomical regions, and distribution shifts, aiming to build holistic evaluation under conditions mirroring clinical deployment.
3. We propose a novel hierarchical prompt framework that leverages structured medical semantics to improve OOD separability. Validated through extensive experiments, our method demonstrates consistent improvements over existing baselines in challenging scenarios.

2 Preliminaries

2.1 Contrastive vision-language models

Recent large vision-language models have shown great potential for various computer vision tasks. In this paper, we focus on CLIP-like models which jointly model the visual and textual data using a dual encoder architecture with one visual encoder $\mathbf{v}_i = f_\theta(\cdot) \in \mathbb{R}^d$ and one text encoder $\mathbf{t}_j = g_\phi(\cdot) \in \mathbb{R}^d$, where θ and ϕ are the corresponding parameters. Formally, for an image x_i out of all images X , the cosine similarity with the specific category y_j out of all candidate prompts Y can be calculated as:

$$s_{i,j} = \frac{\mathbf{v} \cdot \mathbf{t}}{\|\mathbf{v}\| \cdot \|\mathbf{t}\|}. \quad (1)$$

Commonly, we select the text with the highest similarity as the prediction result:

$$\hat{i} = \arg \max_{i=1,\dots,M} \mathbf{s}_i. \quad (2)$$

2.2 OOD Detection with CLIP

A notable advantage of CLIP is its ability, known as zero-shot transfer, to make predictions on any potential categories given a set of candidate prompts. However, this also raises the risk of generating blind predictions for OOD samples that do not belong to any of the provided categories. Given an input x_i , A baseline for OOD detection with CLIP can be formulated as:

$$G(x_i; f, g) = \begin{cases} 1, & S(x_i; f, g) \geq \lambda \\ 0, & S(x_i; f, g) < \lambda \end{cases}, \quad (3)$$

where $S(\cdot)$ is a scoring function to measure the possibility that the input sample x_i is an OOD sample and λ is set manually as the threshold. A simple scoring function can be directly based on the maximum logit score, which is:

$$S_{MS} = -\max \mathbf{s}_i. \quad (4)$$

Under this formulation, the prediction probability with a lower confidence should have a higher probability to be an OOD sample.

2.3 Maximum Concept Matching as Scoring Function

Maximum Concept Matching (MCM) [15] is a state-of-the-art zero-shot OOD detection method that calculates the OOD confidence after softmax function with proper temperature scaling τ . Given the cosine similarity $\mathbf{s}_i$, we have:

$$S_{MCM} = -\max \frac{e^{s_{i,j}/\tau}}{\sum_{j=1}^M e^{s_{i,j}/\tau}}, \quad (5)$$

where the temperature value τ is depended on the downstream datasets. MCM stated that softmax with temperature scaling improves the separability between ID and OOD samples. MCM also suggested that naive maximum softmax probability (MSP) without temperature is suboptimal for zero-shot OOD detection. In this paper, we introduce MCM as a strong baseline for OOD scoring distribution function as its simple nature without the requirements of re-training or complex hyper-parameters tuning. Unless specified, this work uses MCM as the OOD scoring function for all evaluated comparison methods.

3 Methodologies

3.1 Inference with Hierarchical Prompts

Current mainstream medical VLMs commonly employ category names as textual prompts (e.g., "A $\{modality\}$ image showing $\{class\ name\}$ "), an intuitive

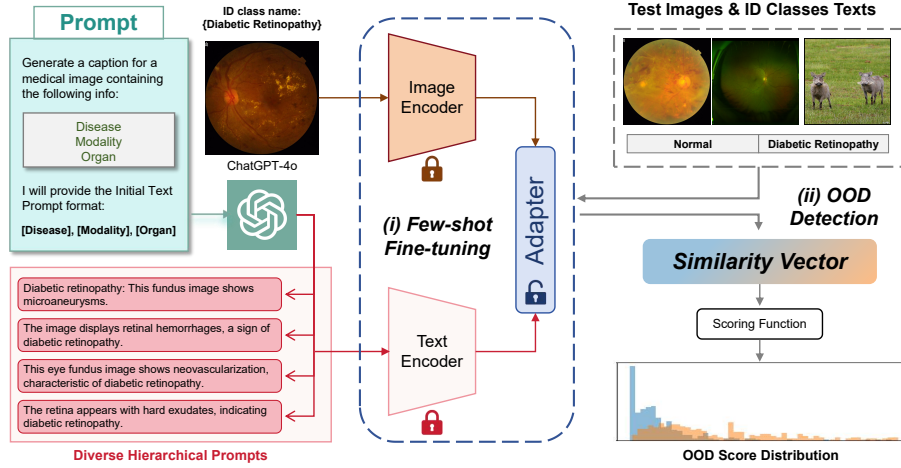


Fig. 2. The fine-tuning pipeline for OOD detection with proposed hierarchical prompts.

approach that achieves reliable performance under standard in-distribution conditions. UniMed-CLIP [9] suggests using diversifying template-captions rather than relying on any single type of prompt as input. Building on this foundation, we develop hierarchical prompts that integrate multi-level clinical semantics, such as diagnostic criteria, lesion morphology, imaging modalities, and anatomical context, to refine the discriminative boundary between ID and OOD samples. Hierarchical prompts are better aligned with medical imaging than natural images due to domain-specific characteristics: (1) Medical diagnosis relies on structured hierarchies (e.g., anatomical location → pathology severity → lesion morphology), which naturally map to multi-level prompts; (2) While natural images often lack standardized descriptors beyond generic labels (e.g., "dog"), medical images require granular, protocol-driven annotations (e.g., "non-proliferative diabetic retinopathy with microaneurysms in the superior quadrant"), enabling prompts to leverage clinically validated taxonomies.

3.2 Few-shot Fine-tuning with Hierarchical Prompts

Given the outstanding generalization abilities of CLIP-like models, it is also evident that fine-tuning plays a crucial role in both ID recognition and OOD detection [20]. Medical CLIP-like models strive to offer a generalized feature representation by pre-training on the collected large-scale image-text pairs [9]. This generalization capacity not only enables effective zero-shot transfer but also boosts the performance of few-shot fine-tuning performance when working with limited downstream data, surpassing the results of training from scratch. Fig. 2 presents an overview of few-shot fine-tuning pipeline, which is also in excellent alignment with our proposed hierarchical prompts [7,8].

Table 1. Evaluated Datasets for Full-spectrum OOD Detection Benchmark.

Foundational Dataset				Covariate Shift OOD Dataset		
Dataset	ID Class Name	Attribute	NoI/NoO	Dataset	Attribute	NoO
FIVES	Normal/DR	Fundus	2 / 2	DeepDRiD	UWF	2
LC25000	Benign/ACA	Lung	2 / 1	LC25000	Colon	2
COVID-19	Normal/Pneu.	X-Ray	2 / 1	COVID-19	CT	2
OOD Class Name ID/OOD Samples				Class Name	OOD Samples	
FIVES	AMD/Glaucoma	300	/ 300	Normal/DR	204	
LC25000	SCC	300	/ 300	Benign/ACA	300	
COVID-19	COVID-19 Pneu.	300	/ 300	Normal/Pneu.	300	

* NoI: Number of ID categories; NoO: Number of OOD categories.

* DR: diabetic retinopathy; AMD: Age-related Macular Degeneration.

* ACA: adenocarcinoma; SCC: Squamous cell carcinomas; Pneu.: Pneumonia.

4 Experiments

4.1 Full-Spectrum Medical OOD Detection Benchmark

Medical Vision-language Models. We include five general-purpose models (GPMs)—Meta CLIP [21], BioMedCLIP [30], PMC-CLIP [12], and UniMed-CLIP [9]—to investigate their generalization capabilities compared to domain-specific medical VLMs. Domain-Specific Models (DSMs) refer to models originally designed for use within a single medical domain. Subsequently, we select one DSM for each medical domain, which are: FLAIR [22] for Ophthalmology; QuilNet [5] for Pathology; and MedCLIP [25] for Radiology.

Datasets. As outlined in Table 1, our benchmark leverages four foundational datasets: FIVES [6], ISIC 2019 [23], LC25000 [1], and COVID-19 [2]. We first define ID categories within each dataset, then construct semantic shifted OOD samples by selecting some other classes that share imaging modalities or anatomical regions with ID data but differ in diagnostic labels. For covariate shifts, we preserve ID semantic labels while introducing distributional variations through differences in imaging devices, acquisition protocols, or population distribution. These covariate shifted OOD samples are sourced both from the foundational datasets and external repositories, such as DeepDRiD [13]. Additionally, we include 300 randomly selected ImageNet samples as far-OOD examples to simulate natural image outliers. ChatGPT-4o was used once to generate hierarchical prompts for each category. Further details, including input prompts used and complete dataset statistics, will be released along with the codes.

Baseline Methods and Setup. To validate the effectiveness of the proposed hierarchical prompts, we evaluate representative zero-shot [14,19,15] and few-shot [31,17,29,27] CLIP-based OOD detection methods with 50-shot fine-tuning. Performance is assessed using the area under the receiver operating characteristic curve (AUROC) for both ID recognition and OOD detection tasks.

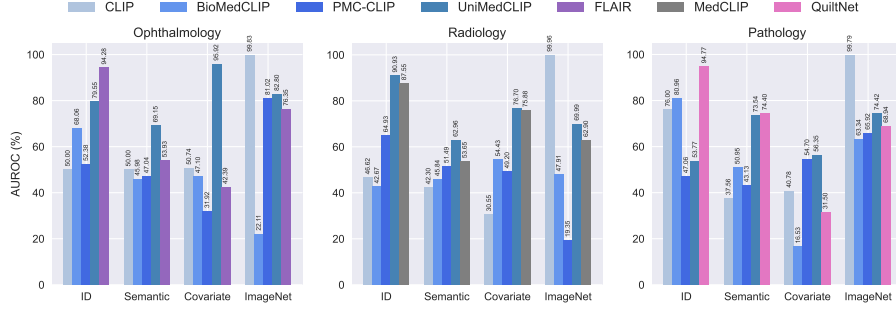


Fig. 3. The comparison results across various GPMs and DSMs.

Table 2. Comparison of representative CLIP-based OOD detection methods.

	Method	FLAIR			UniMedCLIP			QuiltNet		
		S	C	I	S	C	I	S	C	I
<i>Zero-Shot</i>	Max-Logits	40.3	42.8	73.8	44.0	62.5	47.2	60.9	38.1	83.2
	Energy [14]	39.8	43.8	71.4	43.7	58.9	45.4	49.9	43.5	84.0
	GL-MCM [19]	52.1	43.5	74.1	54.5	66.3	54.2	50.3	44.5	55.1
	MCM [15]	53.9	42.4	76.4	53.7	75.9	62.9	55.9	39.8	51.9
	MCM ($L=1$)	61.6	65.6	52.2	67.9	79.5	13.3	48.6	40.3	69.5
	MCM ($L=5/10/3$)	66.7	87.7	82.4	71.1	82.5	35.1	63.0	37.5	59.3
<i>Few-Shot</i>	CoOp [31]	70.3	45.6	90.4	76.4	83.5	56.2	67.2	55.6	66.4
	LoCoOp [17]	72.6	52.3	92.1	74.5	71.2	45.7	55.6	57.2	63.1
	TipAdapter [29]	68.9	50.6	80.4	72.5	80.4	64.5	61.3	40.1	50.0
	HGCLIP [27]	54.5	43.5	56.5	68.3	55.4	50.8	50.4	48.2	55.7
	LoCoOp ($L=5/10/3$)	74.1	62.4	88.3	77.9	82.9	44.2	69.5	49.9	66.7

* S: Semantic shifts; C: Covariate shifts; I: ImageNet;.

4.2 Main Results

Comparison study across various GPMs and DSMs. In Fig. 3, we present the performance differences between GPMs and DSMs in various tasks, including ID recognition and full-spectrum OOD detection. It reveals a critical trade-off: while DSMs achieve stronger ID recognition over GPMs (e.g., 94.28% AUC with FLAIR) through medical fine-tuning, they falter under semantic/covariate OOD shifts, exposing limitations of standard detection methods like MCM. Notably, while Meta CLIP fails to classify medical ID samples but detects far OOD (ImageNet) with 99% AUROC, likely due to its limited medical pre-training enabling learned feature separation from natural images.

Comparison study on advanced methods. We select the models that perform best in ID recognition in each medical domain in Fig. 3, to further examine the performance of advanced CLIP-based OOD detection methods. The results are shown in Table 2. It is found that no single method demonstrates universal superiority across all types of OOD scenarios. We use ($L = 5/3/10$) to denote the levels of hierarchical prompts applied. Specifically, MCM with $L = 5$ shows an

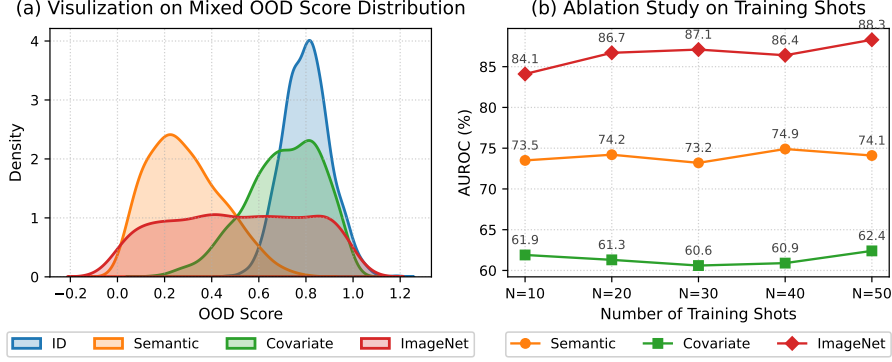


Fig. 4. (a) The visualization on mixed OOD score distribution. (b) Few-shot OOD detection results with different numbers of ID training samples for fine-tuning.

improvement in detecting covariate shifts (e.g., FLAIR achieves 87.7% AUROC) but severely degrades on ImageNet for the other two models. We observe that some medical VLMs tend to classify natural images all into a single category, resulting in the overconfident predictions with catastrophic results (e.g., UniMed-CLIP on MCM ($L=1$) with 13.1% AUROC). This suggests that there should be further potential to improve the scoring functions. Few-shot fine-tuning serves as a universal enhancer for almost all baselines. Relying on the generalization ability of the CLIP-based models, few-shot fine-tuning with ID samples can help focus more on the categories of interest. Meanwhile, hierarchical prompts enrich the models with fine-grained semantic information, thereby enhancing the robustness of the models at critical decision boundaries.

4.3 Analysis

OOD Detection Performance with Mixed Types of Shift. For more practical deployment scenarios, a robust medical VLM must be capable of identifying OOD samples from different types of distribution shifts and reacting accordingly. For instance, when encountering OOD samples with covariate shifts, the system could redirect them to specialized expert models for appropriate ID recognition. In this context, we visualize the score distribution density of FLAIR+LoCoOp ($L=5$) with different OOD types. As shown in Fig. 4-(a), covariate shifts pose a greater challenge due to their higher semantic similarity with ID samples. Meanwhile, the distribution of ImageNet OOD overlaps significantly with that of the other two OOD types. Such misclassifications could potentially trigger unnecessary diagnostic procedures by distributed specialized medical expert models.

Impact of Training Sample Size. Fig. 4-(b) presents the results with different numbers of ID training shots for FLAIR+LoCoOp ($L=5$). Our analysis reveals that LoCoOp demonstrates low sensitivity to the training sample size, achieving robust performance even with minimal ID data. This capability stems from the

strong generalization inherent in pre-trained medical VLMs. Notably, while incorporating additional training samples yields marginal performance gains, the model maintains stable efficacy across all tested configurations.

5 Conclusion

This work establishes the first comprehensive benchmark for medical OOD detection with VLMs, evaluating both general-purpose and domain-specific CLIP-like models under full-spectrum shifts. The proposed hierarchical prompt-based method significantly enhances OOD separability for medical VLMs by leveraging structured medical semantics. We hope that this benchmark and its findings will inspire further research to address critical challenges in VLM-based OOD detection, ultimately contributing to the development of trustworthy and reliable medical diagnostic systems for real-world applications.

6 Acknowledgments

Disclosure of Interests The authors declare no competing interests relevant to the content of this article.

References

1. Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M.: Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142 (2019)
2. Cohen, J.P., Morrison, P., Dao, L.: Covid-19 image data collection. arXiv preprint arXiv:2003.11597 (2020)
3. Graham, M.S., Pinaya, W.H.L., Wright, P., Tudosiu, P.D., Mah, Y.H., Teo, J.T., Jäger, H.R., Werring, D., Nachev, P., Ourselin, S., et al.: Unsupervised 3d out-of-distribution detection with latent diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 446–456. Springer (2023)
4. Hong, Z., Yue, Y., Chen, Y., Cong, L., Lin, H., Luo, Y., Wang, M.H., Wang, W., Xu, J., Yang, X., et al.: Out-of-distribution detection in medical image analysis: A survey. arXiv preprint arXiv:2404.18279 (2024)
5. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems* **36**, 37995–38017 (2023)
6. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific data* **9**(1), 475 (2022)
7. Ju, L., Yu, Z., Wang, L., Zhao, X., Wang, X., Bonnington, P., Ge, Z.: Hierarchical knowledge guided learning for real-world retinal disease recognition. *IEEE Transactions on Medical Imaging* **43**(1), 335–350 (2023)

8. Ju, L., Zhou, Y., Xia, P., Alexander, D., Keane, P.A., Ge, Z.: Explore vision-language model with hierarchical information for multiple retinal disease recognition. *Investigative Ophthalmology & Visual Science* **65**(7), 1593–1593 (2024)
9. Khattak, M.U., Kunhimon, S., Naseer, M., Khan, S., Khan, F.S.: Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372* (2024)
10. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems* **31** (2018)
11. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: *International Conference on Learning Representations* (2018)
12. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 525–536. Springer (2023)
13. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al.: Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns* **3**(6) (2022)
14. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. *Advances in neural information processing systems* **33**, 21464–21475 (2020)
15. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems* **35**, 35087–35102 (2022)
16. Ming, Y., Li, Y.: How does fine-tuning impact out-of-distribution detection for vision-language models? *International Journal of Computer Vision* **132**(2), 596–609 (2024)
17. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems* **36**, 76298–76310 (2023)
18. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. *arXiv preprint arXiv:2304.04521* (2023)
19. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection. *International Journal of Computer Vision* pp. 1–11 (2025)
20. Noda, S., Miyai, A., Yu, Q., Irie, G., Aizawa, K.: A benchmark and evaluation for real-world out-of-distribution detection using vision-language models. *arXiv preprint arXiv:2501.18463* (2025)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
22. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis* **99**, 103357 (2025)
23. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **5**(1), 1–9 (2018)

24. Wang, M., Lin, T., Lin, A., Yu, K., Peng, Y., Wang, L., Chen, C., Zou, K., Liang, H., Chen, M., et al.: Common and rare fundus diseases identification using vision-language foundation model with knowledge of over 400 diseases. arXiv preprint arXiv:2406.09317 (2024)
25. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)
26. Wu, R., Zhang, C., Zhang, J., Zhou, Y., Zhou, T., Fu, H.: Mm-retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 722–732. Springer (2024)
27. Xia, P., Yu, X., Hu, M., Ju, L., Wang, Z., Duan, P., Ge, Z.: Hgclip: exploring vision-language models with graph representations for hierarchical understanding. arXiv preprint arXiv:2311.14064 (2023)
28. Yang, J., Zhou, K., Liu, Z.: Full-spectrum out-of-distribution detection. International Journal of Computer Vision **131**(10), 2607–2622 (2023)
29. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
30. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
31. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)