When Does Uncertainty Become Meaningful in Medical Vision–Language Models? Alignment Enables Disease-wise Uncertainty and Risk-Aware Prediction

Tae Hun Kim[1][0009-0000-8766-4536] and Hyun Gyu Lee[2,⋆][0000-0002-6123-2556]

[1] Department of Electrical and Computer Engineering, Inha University, Incheon, Republic of Korea
ka06052@inha.edu

[2] College of Medicine, Inha University, Incheon, Republic of Korea
hglee@inha.ac.kr

Abstract. Recent Vision–Language Models (VLMs) have shown strong zero-shot performance for chest X-ray interpretation, yet they lack reliable uncertainty estimation, particularly in multi-label clinical settings requiring disease-wise confidence. Existing similarity-based uncertainty methods are largely designed for single-label OOD detection and become unreliable when affirmative and negated disease hypotheses are not semantically aligned, causing uncertainty to reflect representation ambiguity rather than clinical decision uncertainty. We introduce Bi-EDL, a fine-tuning framework that combines bidirectional multiple-choice questioning with evidential deep learning(EDL) to estimate disease-wise uncertainty on aligned image–disease representations. By reinterpreting hypothesis-level similarities as evidence and modeling them as the parameters of a Beta distribution under an evidential deep learning framework, the proposed method mitigates overconfidence inherent in similarity-based scoring and enables pathology-specific uncertainty estimation. Experiments demonstrate improved selective risk controllability under both normal and covariate-perturbed conditions, achieving performance beyond that of the baseline model. Our findings suggest that meaningful uncertainty in medical VLMs emerges when disease hypotheses are semantically aligned, highlighting alignment as a key condition for reliable risk-aware prediction in clinical settings.

Keywords: Uncertainty · Vision-Language Models · Selective Prediction.

1 Introduction

Recent VLMs have demonstrated strong zero-shot performance for chest X-ray (CXR) interpretation, yet clinical deployment additionally requires reliable

⋆ Corresponding author.

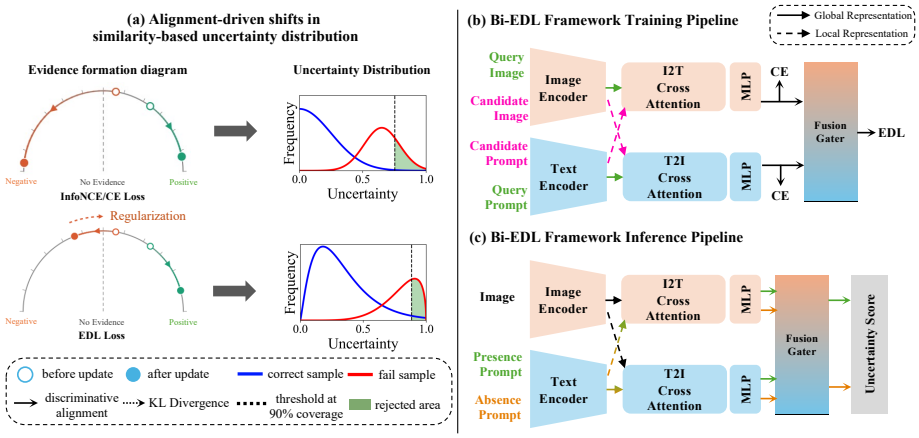

Fig. 1. (a) Evidence formation diagram and resulting uncertainty distributions under InfoNCE/CE and EDL losses. InfoNCE/CE pushes evidence outward without bound, yielding overconfident scores; EDL regularizes incorrect-direction evidence toward Beta(1,1). (b) Bi-EDL training pipeline: bidirectional cross-attention over query–candidate pairs produces I2T and T2I scores, jointly optimized with MCQ cross-entropy and EDL losses. (c) Bi-EDL inference pipeline: fixed presence/absence prompts per disease yield disease-wise uncertainty scores for selective prediction.

uncertainty estimates to support selective automation and human-in-the-loop decision-making. Due to the high computational cost of VLMs, uncertainty estimation in practice has largely relied on similarity-based approaches but most existing methods are designed for single-label OOD detection and fail in multi-label settings where disease-wise confidence is required[1–5]. Moreover, contrastive objectives such as InfoNCE and cross-entropy overly sharpen similarity distributions, inducing overconfident representations that assign high confidence to incorrect predictions, as illustrated in Fig. 1(a).

To address these limitations, we reformulate similarity-based uncertainty estimation for multi-label diagnosis. Specifically, we define the similarities between an image and mutually exclusive disease presence/absence prompts as hypothesis-level evidence, transforming multi-class confidence proxies into disease-wise binary comparisons. To compensate for limited negation understanding in existing VLMs, we apply Bi-directional Multiple Choice Questioning (Bi-MCQ) fine-tuning to align images with affirmative and negated hypotheses along disease-specific semantic axes[6]. Furthermore, we adopt evidential deep learning (EDL) to model the resulting presence/absence similarities as Beta distribution parameters, regularizing overconfident evidence and yielding calibrated disease-wise uncertainty[7].

We refer to this fine-tuning framework as Bi-EDL, whose training and inference pipelines are illustrated in Fig. 1(b) and (c). Bi-EDL infers similarity-based uncertainty scores from disease presence and absence hypotheses to enable disease-wise selective prediction, as shown in Fig. 2. We evaluate whether its

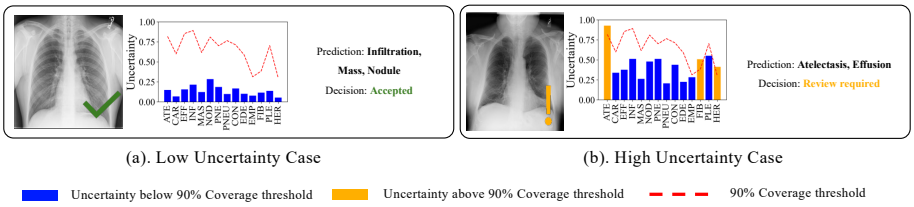

Fig. 2. Disease-wise uncertainty scores under 90% coverage threshold for (a) a low-uncertainty case accepted automatically and (b) a high-uncertainty case referred for human review. Blue bars indicate uncertainty below the threshold; orange bars indicate uncertainty above it.

alignment structure provides practical utility from a coverage–risk perspective, rather than focusing solely on AUROC.

In summary, this work makes three key contributions. (1) We reformulate similarity-based uncertainty estimation for multi-label CXR diagnosis using paired presence/absence hypotheses, enabling systematic reuse of existing similarity-based uncertainty proxies at the disease level. (2) We show that disease-wise hypothesis–image alignment combined with overconfidence regularization via EDL yields a more reliable uncertainty structure. (3) We analyze the resulting uncertainty under selective prediction and demonstrate its practical relevance through improved coverage–risk behavior in clinical settings.

2 Method

We present a framework composed of two coupled components: Bi-MCQ to establish discriminative alignment between images and disease hypotheses, and EDL to suppress overconfident similarity scores on the aligned space. These components are jointly optimized as an integrated framework rather than sequentially applied modules.

2.1 Bi-directional Conditional Alignment via Multiple-Choice Training

To enhance vision–language alignment and improve negation understanding, we adopt Bi-MCQ training based on multi-disease labels. The model jointly learns Image-to-Text (I2T) MCQ, which selects the correct prompt given a query image, and Text-to-Image (T2I) MCQ, which selects the correct image given a query prompt. In the I2T setting, candidate prompts are sampled from affirmative, negative, and mixed statements generated from batch-level annotations, with exactly one correct answer per query. In the T2I setting, candidate images are sampled from the batch, with the single image semantically consistent with the query prompt designated as the correct answer. Here, $y^{\mathrm{I2T}} \in \{1, \ldots, N\}$ denotes the index of the correct prompt among N sampled candidates, and

$y^{\mathrm{T2I}} \in \{1, \ldots, M\}$ denotes the index of the correct image among M sampled candidate images.

In the I2T direction, a global image representation v^g (i.e., the `[CLS]`-level embedding) is used as the query, and cross-attention is applied over the global and local text representations $[t_i^g, t_i^l]$ where t_i^g denotes the `[CLS]`-level text embedding and t_i^l denotes the token-level text features:

$$S_i^{\mathrm{I2T}} = \mathrm{MLP}_{\mathrm{I2T}}(\mathrm{CrossAttn}_{\mathrm{I2T}}(Q = v^g, K = [t_i^g, t_i^l], V = [t_i^g, t_i^l])) \tag{1}$$

The T2I direction is defined symmetrically, where v_j^l denotes the patch-level image features:

$$S_j^{\mathrm{T2I}} = \mathrm{MLP}_{\mathrm{T2I}}(\mathrm{CrossAttn}_{\mathrm{T2I}}(Q = t^g, K = [v_j^g, v_j^l], V = [v_j^g, v_j^l])) \tag{2}$$

Each direction is trained as a multi-class classification problem using cross-entropy:

$$\mathcal{L}_{\mathrm{I2T}} = \mathrm{CE}(S^{\mathrm{I2T}}, y^{\mathrm{I2T}}), \qquad \mathcal{L}_{\mathrm{T2I}} = \mathrm{CE}(S^{\mathrm{T2I}}, y^{\mathrm{T2I}}). \tag{3}$$

The Bi-MCQ objective is defined as

$$\mathcal{L}_{\mathrm{MCQ}} = (\mathcal{L}_{\mathrm{I2T}} + \mathcal{L}_{\mathrm{T2I}})/2. \tag{4}$$

2.2 Disease-Specific Evidence Formulation

In multi-label CXR classification, a single image may be associated with multiple diseases, requiring independent discriminative evidence for each disease. To this end, we construct disease-specific positive and negative prompts and use image–text alignment outputs as the source of disease-wise evidence.

Let $\mathcal{D} = \{d_1, \ldots, d_K\}$ denote the disease set, where K is the number of disease categories. For each disease d_k, we define a positive prompt T_k^+ and a negative prompt T_k^-. $S_{k,\pm}^{\mathrm{I2T}}$ and $S_{k,\pm}^{\mathrm{T2I}}$ denote the I2T and T2I alignment scores for the positive $(+)$ and negative $(-)$ prompts of disease d_k, respectively, and $\tau_{\mathrm{I2T}}, \tau_{\mathrm{T2I}} > 0$ are the corresponding temperature parameters. A learnable Fusion Gater integrates the bidirectional alignment scores to produce disease-specific logits:

$$z_k^{\pm} = w \frac{S_{k,\pm}^{\mathrm{I2T}}}{\tau_{\mathrm{I2T}}} + (1 - w) \frac{S_{k,\pm}^{\mathrm{T2I}}}{\tau_{\mathrm{T2I}}}, \tag{5}$$

where $w \in [0, 1]$ is a learnable weight parameter of the Fusion Gater. We interpret z_k^+ and z_k^- as the sources of evidence supporting the presence and absence hypotheses, respectively. By explicitly defining these opposing inference outputs for each disease, the framework enables disease-wise evidence modeling under the multi-label setting.

2.3 Beta-based Evidential Learning for Disease-wise Calibration

To extend the disease-specific presence and absence logits z_k^+, z_k^- into a distributional evidence framework, we adopt an EDL formulation. Each logit is transformed into non-negative evidence via $e_k^+ = \mathrm{softplus}(z_k^+)$ and $e_k^- = \mathrm{softplus}(z_k^-)$,

and mapped to Beta parameters $\alpha_k^+ = e_k^+ + 1$, $\alpha_k^- = e_k^- + 1$, with total strength $S_k = \alpha_k^+ + \alpha_k^-$. Predictive probabilities are defined as $p_k^+ = \frac{\alpha_k^+}{S_k}$ and $p_k^- = \frac{\alpha_k^-}{S_k}$.

To promote evidence for the correct hypothesis, we employ a matching loss derived from the expected log-likelihood:

$$\mathcal{L}_{\text{match}} = \sum_{k=1}^{K} \left[\psi(S_k) - \psi(\alpha_{k,y}) \right], \tag{6}$$

where $\alpha_{k,y} = y_k \alpha_k^+ + (1 - y_k)\alpha_k^-$.

To suppress excessive incorrect evidence, we introduce a selective KL regularization term. The adjusted parameters are defined as $\tilde{\alpha}_k^+ = y_k + (1 - y_k)\alpha_k^+$ and $\tilde{\alpha}_k^- = (1 - y_k) + y_k \alpha_k^-$. The regularization term is then formulated as

$$\mathcal{L}_{\text{KL}} = \sum_{k=1}^{K} \text{KL}\big(\text{Beta}(\tilde{\alpha}_k^+, \tilde{\alpha}_k^-) \,\|\, \text{Beta}(1, 1)\big). \tag{7}$$

The final evidential objective is

$$\mathcal{L}_{\text{EDL}} = \mathcal{L}_{\text{match}} + \lambda_{KL} \mathcal{L}_{\text{KL}}. \tag{8}$$

Bi-MCQ alignment loss and the evidential loss are jointly optimized. A scheduling coefficient $\lambda_e = \min(1, epoch/epoch_{warmup})$ is applied at epoch to balance the two objectives. This scheduling strategy prioritizes vision–language alignment in early training and gradually increases the contribution of the evidential objective, enabling stable learning of disease-wise evidence strength and uncertainty structure.

$$\mathcal{L}_{\text{total}} = (1 - \lambda_e)\mathcal{L}_{\text{MCQ}} + \lambda_e \mathcal{L}_{\text{EDL}}. \tag{9}$$

Through this formulation, the model increases the evidence supporting the correct hypothesis while suppressing excessive evidence assigned to the incorrect hypothesis. As a result, balanced evidence strength is learned, providing a stable foundation for suppressing overconfident similarity scores during alignment and for reliably estimating similarity-based uncertainty.

3 Experiments

3.1 Dataset, Evaluation Metric and Implementation Details

We fine-tune Bi-EDL on the ChestXray14 dataset[8], which contains 80,718 training, 8,969 validation, and 22,433 test images across 14 disease categories (e.g., Atelectasis, Cardiomegaly, Effusion), exhibiting severe class imbalance ranging from common findings such as Infiltration to rare ones such as Hernia. All classification results are reported on the official test set. Bi-MCQ candidate prompts are constructed solely from the existing disease labels without additional annotation: for each disease d_k, a positive prompt of the form *"There is [disease]"*

Table 1. Comparative analysis of failure detection performance for Bi-MCQ and Bi-EDL across uncertainty methods under Normal and Covariate-perturbations. Better results are highlighted in bold.

		Normal					Covariate-Perturbations				
Metric	**Model**	MSP	ODIN	Energy	MaxLogit	EDL	MSP	ODIN	Energy	MaxLogit	EDL
AUROC(↑)	Bi-MCQ	**0.858**	0.832	**0.855**	**0.855**	0.844	**0.815**	**0.777**	**0.807**	**0.808**	0.503
	Bi-EDL	0.836	**0.850**	0.854	0.854	0.844	0.780	0.765	0.780	0.780	**0.773**
AURC(↓)	Bi-MCQ	0.0275	0.0318	0.0244	0.0247	0.0251	0.0314	0.0376	0.0310	0.0310	0.0670
	Bi-EDL	**0.0173**	**0.0171**	**0.0168**	**0.0168**	**0.0178**	**0.0257**	**0.0282**	**0.0254**	**0.0254**	**0.0260**
Risk@90(↓)	Bi-MCQ	0.0558	0.0637	0.0544	0.0546	0.0562	0.0557	0.0629	0.0541	0.0543	0.0701
	Bi-EDL	**0.0336**	**0.0360**	**0.0335**	**0.0335**	**0.0350**	**0.0362**	**0.0388**	**0.0362**	**0.0362**	**0.0376**

and a negative prompt of the form *"There is no [disease]"* are used (e.g., *"There is Atelectasis"* and *"There is no Atelectasis"*).

CARZero is adopted as the backbone[9], with a ViT-B image encoder and a BioBERT-based text encoder[10, 11]. Since similarity- and logit-based uncertainty baselines require reliable affirmative and negative prompt behavior, comparative uncertainty experiments are conducted using the Bi-MCQ fine-tuned model. Bi-MCQ is fine-tuned on the CARZero backbone using ChestXray14 to ensure consistent representation alignment. For each disease, affirmative and negative prompts are constructed to explicitly assess disease presence versus absence.

Classification performance is evaluated using AUROC. For failure detection, we report AUROC, AURC, Risk@90, and Risk(1). AURC measures risk–coverage trade-offs under selective classification. Risk@90 denotes the error rate at 90% coverage, and Risk(1) corresponds to the full-coverage error rate, reflecting overall model risk. These metrics jointly assess calibration and high-risk sample identification.

Failure detection is evaluated under both normal and covariate-perturbed settings. Covariate-perturbed images are generated using MedMNIST-C corruptions with additional RandomCrop and RandomAffine augmentations[12]. For similarity-based uncertainty estimation, we employed MSP, ODIN, Energy, MaxLogit, and the uncertainty quantification method proposed in EDL[13–16].

Optimization uses Adam with a learning rate of 1×10^{-5} and batch size 64[17]. All experiments are conducted on a single NVIDIA A100 GPU. During training, $\lambda_{KL} = 0.1$ and a 15-epoch warm-up are applied.

3.2 Failure Detection under Normal and Covariate-perturbed Settings

Table 1 compares failure detection performance under Normal and Covariate-perturbed settings using AUROC, AURC, and Risk@90. AUROC measures global ranking consistency between uncertainty and misclassification and can remain high even when some errors persist in low-uncertainty regions, whereas AURC and Risk@90 directly reflect how much risk remains among retained low-uncertainty samples, degrading sharply when misclassified cases are insufficiently separated into high-uncertainty regions.

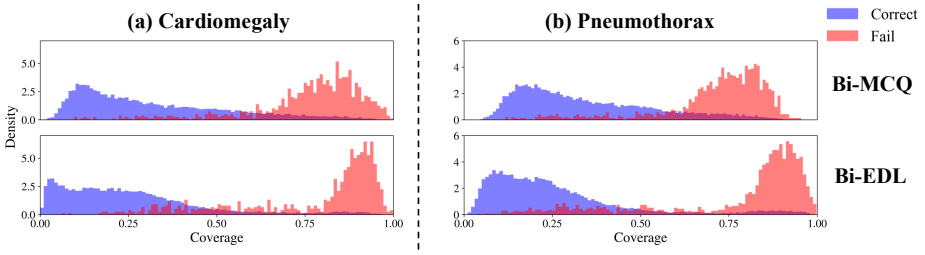

Fig. 3. Histogram distributions of MaxLogit-based uncertainty for correct (blue) and failed (red) samples, comparing Bi-MCQ (top) and Bi-EDL (bottom). Results are shown for (a) Cardiomegaly and (b) Pneumothorax.

Table 2. Ablation results on classification (AUROC) and failure detection with uncertainty-based scoring. Better results are highlighted in bold.

	Classification			Failure Detection		
Model	**Pos AUROC(↑)**	**Neg AUROC(↑)**	**Method**	**AURC(↓)**	**Risk@90(↓)**	**Risk(1)(↓)**
Bi-EDL	**0.854**	**0.854**	MaxLogit	0.0178	0.0350	**0.0549**
			MSP	**0.0173**	**0.0336**	
Bi-MCQ	0.851	0.836	MaxLogit	0.0247	0.0546	0.0796
			MSP	0.0276	0.0558	
EDL Only	0.795	0.795	MaxLogit	0.129	0.164	0.180
			MSP	0.154	0.162	
CARZero	0.811	0.429	MaxLogit	0.229	0.158	0.159
			MSP	0.0633	0.128	

Under the Normal setting, both Bi-MCQ and Bi-EDL achieve comparable AUROC for similarity-based uncertainty scores (e.g., MSP 0.858, MaxLogit 0.855), indicating effective global ranking through discriminative alignment. However, by suppressing overconfident similarities, Bi-EDL consistently reduces selective risk across all uncertainty methods (e.g., MaxLogit: AURC 0.0247 → 0.0168, Risk@90 0.0546 → 0.0335) by preventing misclassified samples from remaining in low-uncertainty regions. Under covariate-perturbed settings, although AUROC degrades due to disruption of the aligned similarity structure, Bi-EDL still achieves lower selective risk than Bi-MCQ (e.g., MaxLogit: AURC 0.0310 → 0.0254, Risk@90 0.0543 → 0.0362). As shown in Fig. 3, Bi-EDL further amplifies uncertainty contrast by lowering uncertainty for correct predictions and increasing it for failures, enabling more effective selective risk control.

3.3 Ablation Study: Alignment and Uncertainty

Table 2 and Fig. 4 present a stepwise analysis of the contributions of alignment (Bi-MCQ) and evidential modeling (EDL). Pos AUROC and Neg AUROC measure discrimination of disease presence and absence. While classification AUROC does not constitute a direct measure of alignment, enhanced separability between presence and absence hypotheses serves as an observable manifestation

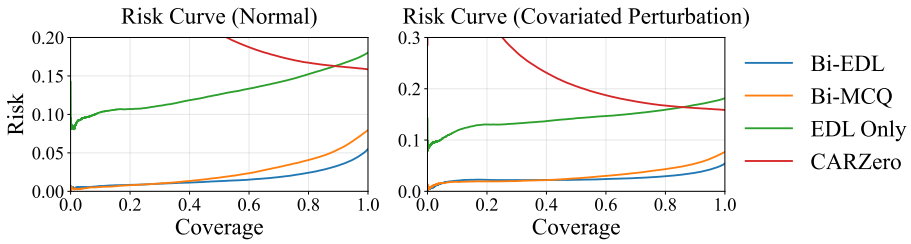

Fig. 4. Risk–coverage curves obtained using Energy-based uncertainty quantification under Normal and Covariate-perturbed environments for different models.

of the disease-wise semantic alignment. Without alignment or uncertainty learning, CARZero shows a low Neg AUROC (0.429) and unstable failure detection performance (AURC 0.229, Risk@90 0.158), suggesting that presence–absence comparison is limited in an unaligned similarity space. Bi-MCQ substantially improves disease-wise discrimination (0.851/0.836) and stabilizes failure detection, providing a structural basis for ranking misclassified samples with higher uncertainty.

EDL Only denotes fine-tuning the CARZero backbone with the EDL objective alone, without disease-wise alignment via Bi-MCQ. While it shows measurable improvements over CARZero, its performance remains below that of Bi-MCQ, indicating that evidential learning alone may not fully substitute for semantic alignment.

Bi-EDL combines evidential learning with alignment, preserving classification performance (0.854/0.854) while further reducing selective risk (e.g., MaxLogit AURC 0.0247 → 0.0178; Risk@90 0.0546 → 0.0335). As shown in Fig. 4, Bi-EDL achieves the lowest risk under both Normal and Covariate-perturbed settings, demonstrating that it mitigates overconfidence and concentrates failure samples in high-uncertainty regions, thereby strengthening selective risk control.

4 Conclusion

This study extends similarity-based uncertainty estimation methods—originally developed under multi-class classification assumptions—to realistic multi-label clinical settings for disease-wise uncertainty estimation in medical VLMs. Specifically, we leverage the similarity between image representations and textual prompts describing disease presence and absence to quantify uncertainty at the disease level. As a fundamental prerequisite for this approach, we identify the importance of semantic alignment between image representations and hypotheses of disease presence or absence. Based on this insight, we propose Bi-EDL, a fine-tuning framework that integrates multiple-choice-based alignment to construct a disease-comparable representation space with EDL to model the relationship between presence and absence hypotheses and mitigate overconfidence in similarity scores.

Experimental results demonstrate that the proposed alignment-aware evidential formulation effectively suppresses overconfident similarity responses and enables stable and interpretable uncertainty quantification in multi-label medical imaging environments. Moreover, the consistent reduction in selective risk indicates that the framework supports selective prediction strategies, enhancing its potential for safe and practical deployment in clinical decision-support settings. This study employs a single medical VLM backbone to isolate the effect of alignment on uncertainty formation rather than to compare architectures. Consistent trends observed across normal and perturbed conditions suggest that the performance gains are associated with alignment-driven representational improvements, while generalization across diverse architectures remains an important direction for future work.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) (RS-2026-25473885) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) (RS-2022-II220641) grants funded by the Ministry of Science and ICT (MSIT), Republic of Korea.

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

References

1. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. In: Advances in Neural Information Processing Systems (NeurIPS), **35**, pp. 35087–35102. (2022)
2. Qin, Y., Peng, D., Peng, X., Wang, X., Hu, P.: Deep evidential learning with noisy correspondence for cross-modal retrieval. In: Proceedings of the ACM International Conference on Multimedia (ACM MM), pp.4948–4956. ACM (2022)
3. Wang, H., Li, Y., Yao, H., Li, X.: CLIPN for zero-shot OOD detection: Teaching CLIP to say no. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1802–1812. (2023)
4. Sun, G., Bai, Y., Yang, X., Fang, Y., Fu, Y., Tao, Z.: Aligning out-of-distribution web images and caption semantics via evidential learning. In: Proceedings of the ACM Web Conference (WWW), pp. 2271-2281. (2024)
5. 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 **133**, 3386–3596 (2025)
6. Kim, T.H., Lee, H.G.: Bi-MCQ: Reformulating vision-language alignment for negation understanding. arXiv preprint arXiv:2601.22696. (2026)
7. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: Advances in Neural Information Processing Systems (NeurIPS), **31** (2018)
8. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2097–2106 (2017)

9. Lai, H., Yao, Q., Jiang, Z., Wang, R., He, Z., Tao, X., Zhou, S.: Carzero: Cross-attention alignment for radiology zero-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11137–11146 (2024)
10. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
11. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
12. Di Salvo, F., Doerrich, S., Ledig, C.: Medmnist-c: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions. MICCAI Workshop on Advancing Data Solutions in Medical Imaging AI (ADSMI @ MICCAI 2024) (2024)
13. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (ICLR) (2017)
14. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations (ICLR) (2018)
15. Liu, W., Wang, X., Owens, J.D., Li, Y.: Energy-based out-of-distribution detection. In: Advances in Neural Information Processing Systems (NeurIPS), **33**, pp. 21464–21475 (2020)
16. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: International Conference on Machine Learning (ICML) (2022)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)