KEPIL: Knowledge-Enhanced Prompt-Image Learning for Prompt-Robust Disease Detection

Haozhe Luo[1,3], Shelley Zixin Shu[1], Ziyu Zhou[2], Robert Berke[3], and Mauricio Reyes[1,4]

[1] ARTORG Center for Biomedical Engineering Research, University of Bern
{haozhe.luo,mauricio.reyes3}@unibe.ch
[2] Shanghai Jiao Tong University
[3] Kaiko.AI, Zurich, Switzerland
[4] Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

Abstract. Vision–language models (VLMs) show promise for clinical decision support in radiology because they enable joint reasoning over radiological images and clinical text, thereby leveraging complementary clinical information. However, radiological findings are long-tailed in practice, leaving some conditions underrepresented and making zero-shot inference essential. Yet current CLIP-style medical VLMs are sensitive to prompt variations and often lack trustworthy external knowledge at inference time, which hinders reliable clinical deployment. We present *KEPIL*, a prompt-robust framework that integrates curated medical knowledge to stabilize zero-shot generalization. KEPIL comprises: (i) *dynamic prompt enrichment* using ontologies with LLM assistance, (ii) a *semantic-aware contrastive loss* aligning embeddings of equivalent prompt variants via a dual-embedding objective, and (iii) *entity-centric report standardization* to yield ontology-aligned representations. Across seven benchmarks, KEPIL achieves state-of-the-art zero-shot inference performance; under prompt-variation tests, it improves AUC by 6.37% on *CheXpert* and by 4.11% on average. These results suggest that structured knowledge and robust prompt design are key to clinically reliable radiology-facing VLMs. Code will be released at https://github.com/Roypic/KEPIL.

Keywords: Chest Xrays · Chest CT · Prompt Robustness · Domain invariance.

1 Introduction

Vision–language models (VLMs)[17,14] have become a compelling paradigm for medical AI by enabling joint reasoning over radiological images and accompanying clinical text. Nevertheless, radiological findings exhibit pronounced long-tailed distributions and impose substantial expert-annotation burdens, leaving many conditions underrepresented during training [26,7,13]. In this context, zero-shot inference is indispensable: for rare diseases, available cases are too limited to

support supervised learning. Notably, standardized medical ontologies and radiographic descriptors are often available even for underrepresented conditions, enabling their semantic alignment with well-characterized common diseases. Such knowledge-driven alignment facilitates reliable diagnosis in data-sparse regimes [27,13,21,30], thereby advancing the practical readiness of VLMs for clinical deployment. Despite this potential, their practical deployment remains limited by two key challenges. First, current CLIP-style medical VLMs [29,9,16,3,24] are highly sensitive to variations in textual prompts. Most existing approaches rely on fixed or template-based prompts during training and inference [27,24,12], yet even minor changes in phrasing at test time can lead to substantial performance degradation as illustrated in Figure 3. This lack of robustness undermines their reliability in real-world clinical settings, where user inputs are naturally diverse. Second, robust generalization to rare or unseen diseases hinges on dependable external knowledge bases. However, prior work often constructs disease knowledge solely via LLMs, inviting hallucination and inconsistency [8,28].

To address these issues, we propose *KEPIL* (Knowledge-Enhanced Prompt Image Learning), a novel framework designed to improve both robustness to prompt variations and transparent zero-shot generalization. We instead ground knowledge in curated resources (e.g., Radiopaedia, UMLS) and use LLMs only to draft and normalize text under ontology constraints. This ontology-aligned, verifiable knowledge base mitigates hallucination and stabilizes zero-shot inference beyond the training distribution. KEPIL incorporates structured medical knowledge through three key components: (1) dynamic prompt enrichment using medical ontologies and large language models to generate diverse, clinically grounded descriptions; (2) a semantic-aware contrastive loss that encourages consistency among embeddings of prompt variants while separating semantically different prompts; and (3) entity-centric report standardization that reduces sensitivity to linguistic variation. Extensive experiments on seven benchmark datasets demonstrate that KEPIL achieves state-of-the-art performance in zero-shot classification and segmentation, with notable improvements in robustness to prompt variation and enhanced recognition of rare conditions. These findings underscore the critical role of robustness to prompt in developing clinically reliable VLMs.

2 Method

Given a dataset $\mathcal{D}$ that comprises N samples, each sample represented by a pair $(\mathbf{x}, \mathbf{t})$, where $\mathbf{x}$ is a radiographic image of size $H \times W$, and $\mathbf{t}$ is its corresponding radiology report. In addition, each sample is accompanied by a collection of definitions and visual depictions for k medically significant entities, which we denote by $\mathbf{d} = \{d_i\}_{i=0}^{k}$. Our goal is to construct a vision-language model $\mathbf{\Phi}$ that, in a zero-shot scenario, computes the probability $\hat{\mathbf{y}}$ indicating the presence of a specific finding in a given test radiograph $\mathbf{x}$ with the detailed description of the finding. To achieve this, we propose a strategy comprising the following components as shown in Figure 1: (1) Collect detailed visual descriptions and general definitions of findings from the knowledge base, (2) Design prompt for

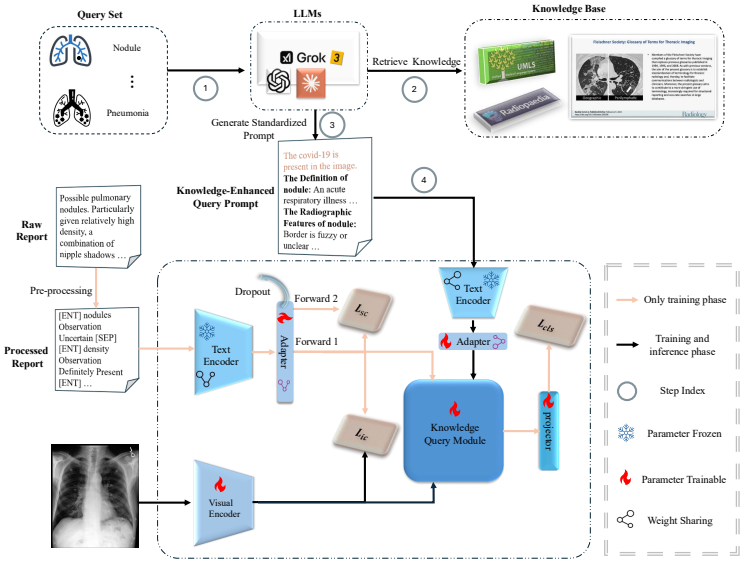

Fig. 1: Overview of the proposed KEPIL framework. The model takes three inputs: a chest X-ray image, its corresponding radiology report, and a knowledge-based entity description. During training, the radiology report is parsed to extract anatomical regions, identified pathologies, and the associated confidence levels of those observations. Two forward passes with different dropout configurations are applied to the adapter to compute the semantic-aware contrastive ($\mathcal{L}_{sc}$). An instance contrastive loss ($\mathcal{L}_{ic}$) is then computed between the report text embeddings and the image embeddings. These embeddings, along with knowledge-based entity embeddings, are fed into a description query network to compute the classification loss ($\mathcal{L}_{cls}$). Knowledge enhancement is achieved by using LLMs to extract definitions and descriptions from medical knowledge bases, producing consistent and structured entity descriptions for training.

pretraining and zero-shot inference, (3) Design effective network architectures and inherit knowledge from other powerful pretraining models, and (4) Devise a training protocol to align the visual and text modalities effectively.

3.1 Knowledge Processing. Following KAD [27], we first mitigate report noise through structured preprocessing. We employ RadGraph [11] to extract radiological entities $\{e_i\}_{i=1}^{j}$ from a report $\mathbf{t}$ and assign each entity a semantic label $s_i \in \{\texttt{ANAT}, \texttt{OBS}\}$ with status (`Definitely Present`, `Definitely Absent`, or `Uncertain`). The preprocessed report is represented as $\mathbf{e} = \{e_1, s_1, \texttt{[SEP]}, \ldots, e_j, s_j\}$, from which we select the top-M most frequent entities to construct an entity set $\mathcal{E}$. To enrich textual supervision beyond standardized medical definitions (e.g., UMLS), we augment entities in $\mathcal{E}$ with structured visual descriptions by

integrating radiographic features from Radiopaedia [20] and the Fleischner glossary [6], and employ large language models to normalize terminology, harmonize format, and complete missing attributes for semantic consistency. Based on the enriched knowledge, we construct a unified prompt template that combines a presence statement (e.g., "<finding> is present in the image"), a concise definition, and characteristic radiographic features (e.g., border clarity, opacity pattern, anatomical location, fluid accumulation, and other salient signs), thereby encoding clinically grounded semantics and promoting tighter alignment between image representations and fine-grained radiological descriptors.

3.2 Network Architecture. Our model adopts a dual-stream design, consisting of an image encoder and a shared text encoder for both the preprocessed report and enriched radiographic descriptions. Given an input chest X-ray $\mathbf{x}$, a visual backbone Φ_{image} produces patch-level features $v = \Phi_{\text{image}}(\mathbf{x}) \in \mathbb{R}^{P\times d}$, where P denotes the number of patches and d the embedding dimension. Under the ViT architecture [5], a learnable `CLS` token $v^{cls} \in \mathbb{R}^d$ summarizes the global image representation. For textual input, the structured report $\mathbf{e}$ and its corresponding enriched prompts $\{p_i\}_{i=0}^{k}$ are encoded via a shared text encoder Φ_{text}, yielding token-level embeddings $t_{e^*} = \Phi_{\text{text}}(\mathbf{e}) \in \mathbb{R}^{T\times d}$ and $t_{p^*} = \Phi_{\text{text}}(p_i) \in \mathbb{R}^{T\times d}$, where T is the token length. The encoder also produces global representations $t_{e^*}^{cls}$ and $t_{p^*}^{cls}$ to capture holistic semantics, and lightweight adapters further refine them into t_e and t_p. To enhance fine-grained diagnostic reasoning, we introduce a Knowledge Query Module (KQM) that performs token-level cross-attention between visual patches and textual knowledge. Specifically, prompt embeddings $t_p^{cls} \in \mathbb{R}^{S\times d}$ are projected to queries Q, while patch-wise visual embeddings $v \in \mathbb{R}^{P\times d}$ are projected to keys K and values V. The knowledge-enhanced representation is obtained via scaled dot-product attention, $t_{attn}^{cls} = \text{softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)V$, where $t_{attn}^{cls} \in \mathbb{R}^{S\times d}$. The fused representations are subsequently mapped to logits and transformed into diagnostic probabilities. Figure 1 provides an overview of the overall architecture.

3.3 Training Objectives. (i) ***Semantic-aware contrastive learning.*** To enhance the expressive power of the frozen text encoder without modifying its pretrained weights, we introduce an adapter module. This module aims to improve the discriminative capability of raw text embeddings and is trained using a contrastive loss ($\mathcal{L}_{sc}$) that maximizes consistency between the original and adapter-enhanced representations. Specifically, for a radiograph-report pair $(\mathbf{x}, \mathbf{t})$, we first extract the frozen text embedding from the `CLS` token, denoted as $t_{e^*}^{cls}$. The adapter then transforms this embedding as follows: $t_e^{cls} = \mathcal{A}(t_{e^*}^{cls})$, where $\mathcal{A}$ denotes the adapter function. We repeat this process to obtain $t_e^{cls^2}$. Due to the dropout in the adapter, t_e^{cls} and $t_e^{cls^2}$ are two perturbed views. The adapter is utilized to learn the consistency from two views of the embedding and the contrastive loss $\mathcal{L}_{sc}$ is applied to maximize the semantic consistency between t_e^{cls} and $t_e^{cls^2}$ while ensuring they are discriminated against embeddings from other samples in a training batch B. Formally, the loss is defined

as: $\mathcal{L}_{sc}(t_e^{cls}, t_e^{cls^2}) = -\log \frac{\exp\left(t_e^{cls}, t_e^{cls^2})/\tau\right)}{\sum_{t_e'^{cls} \in B} \exp\left(\mathrm{s}(t_e^{cls}, t_e'^{cls})/\tau\right)}$. where $\mathrm{s}(\cdot,\cdot)$ represents the cosine similarity and τ is the temperature parameter. B represents the batch.

(ii) ***Instance level alignment of radiograph-report.*** The global alignment between an image and its corresponding radiology report is accomplished using the standard image-text contrastive loss, denoted as $\mathcal{L}_{ic}$. This loss function maximizes the mutual information between the image and report representations. In our approach, we specifically utilize the report embedding because it effectively captures the overall context of the image. Given a triplet consisting of a radiograph, its report, and a description $(\mathbf{x}, \mathbf{e}, \mathbf{d})$, our goal is to enhance the similarity between the embedded image and report features, with a particular focus on their `CLS` token representations. Formally, within a training batch B, the loss is defined as: $\mathcal{L}_{ic}(v^{cls}, t_e^{cls}) = -\log \frac{\exp\left(\mathrm{s}(v^{cls}, t_e^{cls})/\tau\right)}{\sum_{t_e'^{cls} \in B} \exp\left(\mathrm{s}(v^{cls}, t_e'^{cls})/\tau\right)}$.

(iii) Overall loss. The overall training objective integrates the contrastive report embedding loss, the instance-level image-text contrastive loss (Eq. 2) and the binary classification loss. Additionally, we employ a binary cross-entropy loss ($\mathcal{L}_{cls}$), which is applied only when queries from $\mathcal{E}$ explicitly appear or are explicitly absent in the report. If the presence of findings is uncertain, this loss is not computed. The total loss to be minimized is: $\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{ic} + \lambda_3 \mathcal{L}_{sc}$, where λ regulates the contribution of radiographic descriptions.

3 Experiment Settings

We evaluate our approach on seven publicly available chest-X-ray datasets for zero-shot inference: CheXpert [10], ChestXray-14 [23], PadChest [4], RSNA Pneumonia [25], SIIM-ACR [1], COVIDx CXR-2 [18] and LIDC-IDRI dataset [2]; The text encoder is initialized with BioClinicalMPBERT [12] weights during pretraining and remains frozen. The image encoder is ViT-B/16 which utilizes M3AE for pretraining only on the MIMIC dataset [12]. For the adapters, we use light-weighted two-layer MLPs with dropout set to 0.5 by default at the report branch. The adapter, projector, visual encoder and KQM module are trainable. The ChatGPT-4o is utilized for prompt standardization and completion as it performs the best as shown in Table 3. For generating prompt variants, we include not only rephrasings but also realistic errors such as typos, omissions, and incorrect punctuation. The pretraining is conducted on two H100 GPUs.

4 Results

(i) ***Classification for seen diseases.*** Table 1 presents the zero-shot classification results for *seen* diseases. *Seen* diseases are those that appear in the MIMIC dataset, which the model is exposed to during pretraining. Compared to other vision-language models (VLMs) pretrained on the MIMIC dataset, our proposed method, KEPIL, consistently achieves the highest performance in terms of both

Table 1: Comprehensive zero-shot evaluation on seen and unseen diseases. Seen datasets contain diseases observed during pretraining. Unseen settings include novel categories, rare diseases, and modality shift (CXR→CT). Best results are bold; second-best are underlined.

Method	Seen Diseases (CXR)															Unseen / Rare / Modality Shift						
	CheXpert			ChestXray-14			PadChest-seen			RSNA			SIIM			Covid		PadChest-unseen		PadChest-rare		LIDC
	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
ConVIRT [29]	52.10	35.61	57.43	53.15	12.38	57.88	63.72	14.56	73.47	79.21	55.67	75.08	64.25	42.87	53.42	62.78	63.84	51.17	61.51	50.37	60.17	-
GLoRIA [9]	54.84	37.86	60.70	55.92	14.20	59.47	64.09	14.83	73.86	70.37	48.19	70.54	54.71	40.39	47.15	64.52	60.21	49.96	60.95	48.25	58.49	-
BioViL [3]	60.01	42.10	66.13	57.82	15.64	61.33	60.35	10.63	70.48	84.12	54.59	74.43	70.28	46.45	68.22	61.40	58.20	57.95	62.50	52.82	60.60	-
BioViL-T [3]	70.93	47.21	69.96	60.43	17.29	62.12	65.78	15.37	77.52	86.03	62.56	80.04	75.56	60.18	73.72	62.43	57.65	58.94	68.56	57.44	65.38	-
CheXzero [22]	87.90	61.90	81.17	66.99	19.95	65.38	73.24	19.53	83.49	83.13	61.49	78.34	84.60	65.97	77.34	73.13	71.45	66.70	81.19	65.08	81.17	-
MedKLIP [24]	87.97	63.67	84.32	72.33	24.18	79.40	77.87	26.63	92.44	86.57	63.28	79.97	89.79	72.73	83.99	76.28	71.96	60.31	76.69	59.75	77.84	45.56
KAD [27]	89.23	63.25	86.25	76.49	29.98	80.49	80.94	58.23	92.12	85.32	87.09	81.80	87.39	68.84	81.73	74.03	72.42	73.08	83.23	72.21	86.54	57.75
MAVL [19]	90.13	65.47	86.44	73.57	26.25	82.77	78.79	28.48	92.56	86.91	63.41	82.42	92.04	77.95	87.14	**83.86**	78.07	70.42	84.00	70.06	84.64	41.81
CARZero [12]	**92.38**	40.20	86.93	79.64	31.41	79.60	83.97	25.08	91.46	80.28	28.28	41.52	90.90	75.96	85.03	75.53	68.55	78.95	85.65	77.79	91.77	60.57
DeViDe [15]	89.87	62.88	87.98	77.61	31.40	82.06	84.26	58.29	92.16	88.58	88.40	84.00	89.50	72.24	83.72	73.75	72.34	75.06	90.84	73.73	91.09	47.90
KEPIL (Ours)	91.21	**65.56**	**87.88**	**80.95**	**34.74**	**83.23**	**85.32**	**59.79**	**93.16**	**89.76**	**90.23**	**86.24**	**93.02**	**79.22**	**87.46**	79.55	**78.12**	**79.05**	**94.02**	**78.75**	**95.47**	**66.65**

AUC and ACC across all five datasets. For instance, KEPIL outperforms the second-best method, CARZero, by 1.31% on the ChestXray-14 dataset and by 2.85% on the RSNA Pneumonia dataset in terms of AUC. Additionally, KEPIL achieves the best or second-best performance on F1 scores. These results demonstrate the model's robustness and effectiveness in zero-shot disease detection under significant domain shifts.

(ii) ***Classification for unseen and rare diseases.*** Table 1 reports zero-shot classification on diseases absent from pretraining ("unseen") and on diseases sparsely observed during pretraining ("rare"). On the two unseen datasets, COVID-19 CXR-2 and PadChest-unseen, our proposed method, KEPIL, achieves either the best or second-best performance when compared to other vision-language models. Specifically, KEPIL demonstrates comparable performance to MAVL on the COVID-19 CXR-2 dataset and achieves notable improvements of 8.37% in ACC on the PadChest-unseen dataset over CARZero. For rare diseases, which are included during pretraining but with very few positive sample counts, KEPIL significantly improves ACC by 3.70% over the second-best performances. These results underscore KEPIL's strong capability to generalize to novel or low-resource disease categories, further highlighting its potential in real-world clinical applications where data scarcity and domain shifts are common. Moreover, we further evaluated KEPIL's cross-modality generalization on CT slices from the LIDC-IDRI dataset, where it achieved a leading AUC performance that surpassed the second-best method by 6.08%.

(iii) ***Robustness to Semantic Variations of Prompts.*** Figure 3 reports zero-shot AUCs using query prompts generated by Grok3, ChatGPT-4o, and Claude Sonnet 4. Across both in-domain and out-of-domain datasets, KEPIL consistently surpasses other SOTA VLMs, demonstrating strong robustness to prompt variability—including realistic errors such as typos, omissions, and incorrect punctuation/formatting. Leveraging this robustness, KEPIL operates with minimal specification: given only a finding name, it automatically composes comprehensive, clinically grounded queries, thereby streamlining prompt creation and improving diagnostic accuracy, transparency, and efficiency. In Table 4, we

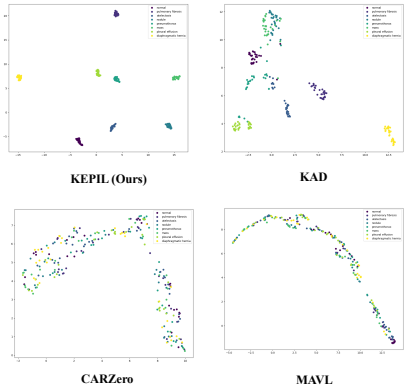

Fig. 2: UMAP projection of embeddings generated by multiple clinical text encoders from 30 prompt variations per disease term. KEPIL exhibits tighter clusters, demonstrating superior robustness to prompt variations.

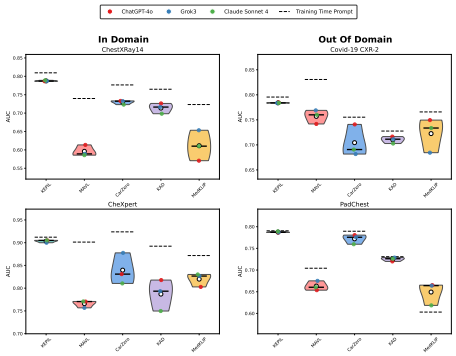

Fig. 3: Evaluation of the impact of prompts from ChatGPT-4o, Grok-3, and Claude Sonnet 4 (including syntactic paraphrases, typos, omissions, punctuation variants) on chest X-ray model performance relative to the original training prompt. KEPIL demonstrates smaller performance declines.

compare our $\mathcal{L}_{sc}$ with naive text augmentation. our $\mathcal{L}_{sc}$ objective stays strictly within the clinical language manifold and enforces invariance without distorting medical semantics, leading to higher robustness across all real clinical variants and better OOD performance.

(iv) ***Richer Textual Inputs Enhance Diagnostic Performance.*** We investigated the impact of varying levels of textual detail on KEPIL's performance. As shown in Figure 4, model performance consistently improves with richer textual inputs. Specifically, using only finding names (e.g., "Atelectasis") results in the lowest AUC scores, while supplementing the names with definitions from the UMLS knowledge base leads to notable gains. The highest performance is achieved when definitions are further enriched with detailed descriptions, highlighting the model's ability to establish a connection between diagnosis and finding-related features. This finding underscores the importance of incorporating rich and structured textual context to enhance the model's robustness and classification accuracy.

(v) ***Enriched Prompt and $\mathcal{L}_{sc}$ trigger Significant Performance Gains.*** We conduct an ablation study to assess the contribution of each component in the KEPIL framework by progressively adding the Enriched Prompt (EP) and $\mathcal{L}_{sc}$ to the base optimization strategy ($\mathcal{L}_{cls} + \mathcal{L}_{ic}$). As shown in Table 2, incorporating enriched prompts significantly improves zero-shot classification performance, boosting the AUC from 76.89% to 79.45% on ChestXray14 and from 87.07% to 90.47% on CheXpert. Adding $\mathcal{L}_{sc}$ with a dropout rate of 0.5 further enhances performance, particularly on ChestXray14, reaching the highest AUC of 80.95%.

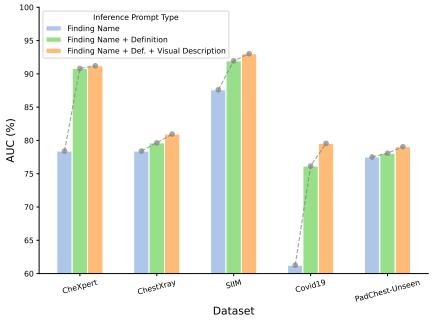

Fig. 4: KEPIL's performance across incremental different inference input types. Performance improves with increasingly rich textual input, highlighting the value of more informative prompts.

Table 2: Ablation study of the effects of enriched prompts, the inclusion and placement of $\mathcal{L}_{sc}$, and dropout rate on zero-shot inference performance.

Ablation	ChestXRay14	CheXpert
Effect of Enriched Prompt and $\mathcal{L}_{sc}$		
Base ($\mathcal{L}_{cls}$ + $\mathcal{L}_{ic}$)	76.89	87.07
+ Enriched Prompt (EP)	79.45	90.47
+ EP + $\mathcal{L}_{sc}$	**80.95**	**91.21**
Effect of Dropout Rate in $\mathcal{L}_{sc}$		
Dropout = 0.3	80.11	90.98
Dropout = 0.4	80.29	91.41
Dropout = 0.5	**80.95**	**91.21**
Dropout = 0.6	80.24	91.24
Effect of the Position of $\mathcal{L}_{sc}$		
Report branch	**80.95**	**91.21**
Prompt branch	79.72	90.52
Report branch + Prompt branch	80.19	91.15

Table 3: Evaluation of LLM generated descriptions against ground-truth.

Metric	ChatGPT-4o	Mixtral-8×7B	Medorca-2x7B	PMC-llama-7B
BLEU-4	**0.1180**	0.1030	0.0018	0.0033
ROUGE-L	**0.2695**	0.2328	0.1461	0.1460
F1 RadGraph	**0.2384**	0.2046	0.0639	0.0979

Table 4: Comparison of $\mathcal{L}_{sc}$ and text side augmentation's impact on zero-shot inference. $\mathcal{L}_{sc}$ remains strictly within the clinical language manifold, yielding real-world robustness. Achiving at most 23.96% improvement compare to CARZero. Numbers in parentheses indicate absolute improvement over CARZero.

Pretraining Setting	*In-domain*		*OOD*		*Prompt Variants (on CheXpert)*			
	CheXpert	CXR14	PadChest	COVID	Abbreviation	Multilingual	Clinician Jargon	Medical Synonym
CARZero	-		-	-	78.81	62.07	79.85	85.21
KEPIL+ $\mathcal{L}_{sc}$	91.21	**80.95**	**79.05**	**79.55**	89.54 (↑10.73)	86.03 (↑23.96)	**89.89 (↑10.04)**	**90.62 (↑5.41)**
+ Text augmentation	**91.78**	80.87	77.28	75.57	**89.63**	**87.84**	83.15	89.87

Experiments with dropout rates between 0.3 and 0.6 show 0.5 to be optimal. We also evaluated where to apply $\mathcal{L}_{sc}$ and found that placing it in the report branch yields the greatest improvement. Overall, the complete KEPIL model achieves a performance gain of over 4% in AUC compared to the base setup across datasets, demonstrating the effectiveness of enriched prompts and the proposed $\mathcal{L}_{sc}$.

5 Conclusion

We introduce KEPIL, a robust knowledge-enhanced vision-language framework designed to address the critical limitations of prompt sensitivity and poor generalization in open-vocabulary disease detection for radiological imaging. Unlike prior VLMs that rely on fixed or template-based prompts, KEPIL integrates medical ontologies, radiographic descriptors, and entity-level representations into the vision-language pipeline. KEPIL establishes a semantically aligned and clinically grounded interface between text and image modalities with a semantic-

aware contrastive learning objective and a prompt standardization protocol, leading to enhanced stability under prompt perturbations and improved generalization to rare and unseen diseases.

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

References

1. Anna Zawacki, C.W., George Shih, J.E., Mikhail Fomitchev, Mohannad Hussain, P., Phil Culliton, S.B.: Siim-acr pneumothorax segmentation. `https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation` (2019)
2. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
3. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
4. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Muller, N.L., Remy, J.: Fleischner society: glossary of terms for thoracic imaging. Radiology **246**(3), 697–722 (2008)
7. Holste, G., Wang, S., Jiang, Z., Shen, T.C., Shih, G., Summers, R.M., Peng, Y., Wang, Z.: Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In: MICCAI Workshop on Data Augmentation, Labelling, and Imperfections. pp. 22–32. Springer (2022)
8. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)
9. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)
10. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)

11. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
12. Lai, H., Yao, Q., Jiang, Z., Wang, R., He, Z., Tao, X., Zhou, S.K.: Carzero: Cross-attention alignment for radiology zero-shot classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11137–11146 (2024)
13. Lin, M., Holste, G., Wang, S., Zhou, Y., Wei, Y., Banerjee, I., Chen, P., Dai, T., Du, Y., Dvornek, N.C., et al.: Cxr-lt 2024: A miccai challenge on long-tailed, multi-label, and zero-shot disease classification from chest x-ray. Medical Image Analysis p. 103739 (2025)
14. Luo, H., Shu, S.Z., Zhou, Z., Otalora, S., Reyes, M.: Xbench: A comprehensive benchmark for visual-language explanations in chest radiography. arXiv preprint arXiv:2510.19599 (2025)
15. Luo, H., Zhou, Z., Hou, M., Royer, C., Reyes, M., Sekuboyina, A., Menze, B.: Devide: Faceted medical knowledge to enhance vision foundation model pretraining for radiology. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 1779–1782. IEEE (2025)
16. Luo, H., Zhou, Z., Shu, S.Z., Mortanges, A.P.d., Berke, R., Reyes, M.: On the interplay of human-ai alignment, fairness, and performance trade-offs in medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 420–430. Springer (2025)
17. Pahud de Mortanges, A., Luo, H., Shu, S.Z., Kamath, A., Suter, Y., Shelan, M., Pöllinger, A., Reyes, M.: Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. NPJ digital medicine **7**(1), 195 (2024)
18. Pavlova, M., Terhljan, N., Chung, A.G., Zhao, A., Surana, S., Aboutalebi, H., Gunraj, H., Sabri, A., Alaref, A., Wong, A.: Covid-net cxr-2: An enhanced deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Frontiers in Medicine **9**, 861680 (2022)
19. Phan, V.M.H., Xie, Y., Qi, Y., Liu, L., Liu, L., Zhang, B., Liao, Z., Wu, Q., To, M.S., Verjans, J.W.: Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11492–11501 (2024)
20. Radiopaedia: `https://radiopaedia.org/` (2024)
21. Rahman, U., Basu, A., Khattak, M.U., Rahman, A.U.: Xdt-cxr: Investigating cross-disease transferability in zero-shot binary classification of chest x-rays. arXiv preprint arXiv:2408.11493 (2024)
22. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)
23. 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. pp. 2097–2106 (2017)
24. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21372–21383 (2023)

25. Wu, L., Zhang, J., Wang, Y., Ding, R., Cao, Y., Liu, G., Liufu, C., Xie, B., Kang, S., Liu, R., et al.: Pneumonia detection based on rsna dataset and anchor-free deep learning detector. Scientific Reports **14**(1), 1929 (2024)
26. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical image analysis **91**, 102996 (2024)
27. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)
28. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al.: Siren's song in the ai ocean: A survey on hallucination in large language models. Computational Linguistics pp. 1–46 (2025)
29. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
30. Zhang, Z., Yu, Y., Chen, Y., Yang, X., Yeo, S.Y.: Medunifier: Unifying vision-and-language pre-training on medical data with vision generation task using discrete visual representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29744–29755 (2025)