Knowledge-Enhanced Explainable Prompting for Vision-Language Models

Yequan Bie[1], **Andong Tan**[1], **Zhixuan Chen**[1], **Zhiyuan Cai**[1], **Luyang Luo**[2], **Hao Chen**[1*]

[1]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
[2]Department of Biomedical Informatics, Harvard University
ybie@connect.ust.hk

Abstract

Large-scale vision-language models (VLMs) embedded with expansive representations and visual concepts have showcased significant potential in image and text understanding. Efficiently adapting VLMs such as CLIP to downstream tasks like few-shot image classification has garnered growing attention, with prompt learning emerging as a representative approach. However, most existing prompt-based adaptation methods, which rely solely on coarse-grained textual prompts, suffer from limited performance and interpretability when handling domain tasks that require specific knowledge. This results in a failure to satisfy the stringent trustworthiness requirements of Explainable Artificial Intelligence (XAI) in high-risk scenarios like healthcare. To address this issue, we propose a **K**nowledge-**E**nhanced **E**xplainable **P**rompting (**KEEP**) framework that leverages fine-grained domain-specific knowledge to enhance the adaptation process of VLMs across various domains and image modalities. By incorporating retrieval augmented generation and domain foundation models, our framework can provide more reliable image-wise knowledge for prompt learning in various domains, alleviating the lack of fine-grained annotations, while offering both visual and textual explanations. Extensive experiments and explainability analyses conducted on eight datasets of different domains and image modalities demonstrate that our method simultaneously achieves superior performance and interpretability, highlighting the effectiveness of the collaboration between foundation models and XAI.

Code — https://github.com/Tommy-Bie/KEEP

1 Introduction

Recent studies in large-scale pre-trained VLMs, such as CLIP (Radford et al. 2021), BLIP (Li et al. 2022), ALIGN (Jia et al. 2021), and Coca (Yu et al. 2022) have highlighted the potential of foundation models (FMs) in vision and language understanding. The effectiveness of large-scale image-text pairs and their alignment has been demonstrated in enhancing vision-language models, enabling them to excel in various tasks (Zhang et al. 2024a). However, the massive sizes and high training costs have prompted researchers to explore efficient methods for adapting the pre-trained VLMs to downstream tasks.

*Corresponding author.

Recently, prompt learning (Zhou et al. 2022a,b), introduced from the field of natural language processing, has emerged as one of the representative methods for efficiently adapting FMs to downstream tasks like image classification. These methods focus on learning the prompts instead of training the whole model, achieving both promising performance and much lower training cost. Traditional prompt learning methods only use one general sentence as the input prompt (e.g., *a photo of a* [*class name*]) (Zhou et al. 2022b; Gao et al. 2021; Khattak et al. 2023), which exhibits relatively low performance when handling fine-grained tasks. Some studies tend to alleviate this issue by introducing knowledge into prompt learning (Yao et al. 2023; Bulat et al. 2023; Kan et al. 2023). However, most existing knowledge-related methods use only coarse-grained textual prompts (e.g., class-level prompts without fine-grained knowledge). This leads them to perform well in some natural image tasks but still exhibit limited performance in various domains due to the lack of domain knowledge. The coarse-grained and insufficient information embedded in these models leads to unsatisfactory interpretability and cannot meet the trustworthiness requirements of XAI, especially in high-stakes scenarios such as healthcare (Hou et al. 2024; Nie et al. 2025).

To address the above issues, we propose **KEEP**, a knowledge-enhanced explainable prompting framework that incorporates the fine-grained knowledge priors elicited from domain-specific FMs to enhance the adaptation of VLMs. As shown in Figure 1, unlike current methods that can only perform well in certain areas, our method unifies the prompt creation and prompt learning process for different domains and image modalities, making full use of domain-specific knowledge to handle various datasets while providing both visual and textual explanations to improve trustworthiness.

We summarize our main contributions as follows: (i) We propose a knowledge-enhanced explainable prompting framework that leverages fine-grained domain-specific knowledge elicited from foundation models and RAG to enhance the VLM adaptation. A knowledge-aware attention module is further adopted to learn and align the semantic correspondences between images and knowledge-enhanced prompts. (ii) We demonstrate that our method can be effectively and flexibly applied to various domains and different image modalities from medical and natural fields. (iii) Extensive experiments and explainability analyses show that

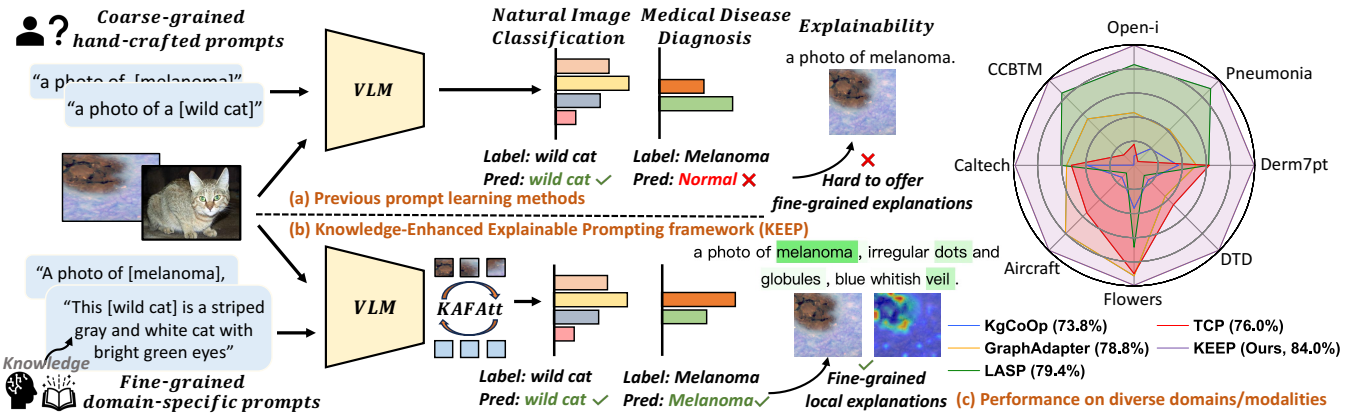

Figure 1: Illustration of Knowledge-Enhanced Explainable Prompting framework (**KEEP**) for various domains: (a) Previous works adopt only coarse-grained general prompts and usually perform well in limited domains. (b) **KEEP** utilizes domain knowledge-enhanced prompts to facilitate bridging the domain gap while offering fine-grained explanations. (c) Performance comparison with state-of-the-art methods on a diverse set of domains and image modalities.

our method concurrently achieves promising performance and interpretability. Our method consistently outperforms general and domain-specific methods with promising data efficiency, while being more explainable by offering visual and textual explanations, highlighting the effectiveness of the collaboration between FMs and XAI.

2 Related Work

2.1 Foundational Vision-Language Models

Vision-language models (VLMs) such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021) and Coca (Yu et al. 2022), are a fusion of vision and natural language models trained on large-scale multi-modal datasets. Among existing VLMs, CLIP is one of the most representative and commonly used frameworks aligning the feature spaces of vision and text encoder via contrastive learning based on 400 million image-text pairs. Recently, the application of large-scale pre-trained VLMs in other domains such as healthcare attracts increasing attention. These domain-specific FMs aim to introduce the vision-language learning approach to medical vision and text understanding. For example, KAD (Zhang et al. 2023b) introduces knowledge graphs with medical concepts into contrastive learning between radiological images and reports. In this work, we elicit fine-grained knowledge from domain-specific foundation models to handle tasks of different image modalities and domains.

2.2 Prompt Learning

In order to address the challenge of the high computational cost of fully fine-tuning VLMs such as CLIP to downstream tasks, prompt learning techniques (Gu et al. 2023; Zhou et al. 2022a,b; Yu et al. 2023) have been introduced as efficient and effective adaptation methods from the field of natural language processing (Liu et al. 2023). Prompt learning aims to improve VLM adaption by inferring a set of learnable textual tokens combined with the class tokens instead of fixing the input textual prompt such as the hand-crafted template of CLIP (*a photo of a* [*class name*]). For instance, CoOp (Zhou et al. 2022b) proposes to replace the fixed hand-crafted prompts with soft/learnable prompts and optimize the textual tokens. CoCoOp (Zhou et al. 2022a) extends CoOp by proposing image-conditional prompts fusing the visual features and the textual prompts. However, these methods with only one simple and global sentence as the input prompt show low performance when handling fine-grained tasks. Some recent studies, e.g., KgCoOp (Yao et al. 2023), LASP (Bulat et al. 2023), TCP (Yao et al. 2024), introduce knowledge to optimize context using more class-level textual templates, which still exhibit limited performance in specific domains due to the lack of domain knowledge such as clinical knowledge. To alleviate the issue, some recent approaches (Kan et al. 2023; Cao et al. 2024; Tan et al. 2024; Bie et al. 2024b; Lee et al. 2025; Koleilat et al. 2025) integrate external knowledge from foundation models, e.g., CoAPT (Lee et al. 2025) and BiomedCoOp (Koleilat et al. 2025) utilizes foundation models to generate textual knowledge such as class attribute words, which we argue is still coarse-grained due to the usage of only category-level knowledge, leading to suboptimal performance and lack of explainability. Therefore, we propose leveraging image-wise domain-specific knowledge to enhance the adaptation process, while improving model interpretability by providing prompt-based explanations from different perspectives.

2.3 Knowledge-based XAI

Bridging the understandability gap between humans and black-box AI models requires techniques that can answer the multifaceted problem of explainability, addressing the faithfulness (Lakkaraju et al. 2019) of the explanations representing the model's behavior, while also considering the capability of the human interpreter to understand it. Domain-specific knowledge, derived from human knowledge in various fields, plays an important role in improving the model performance and explainability (Tocchetti et al. 2022; Bie et al. 2024a). For example, Concept Transformer (Rigotti

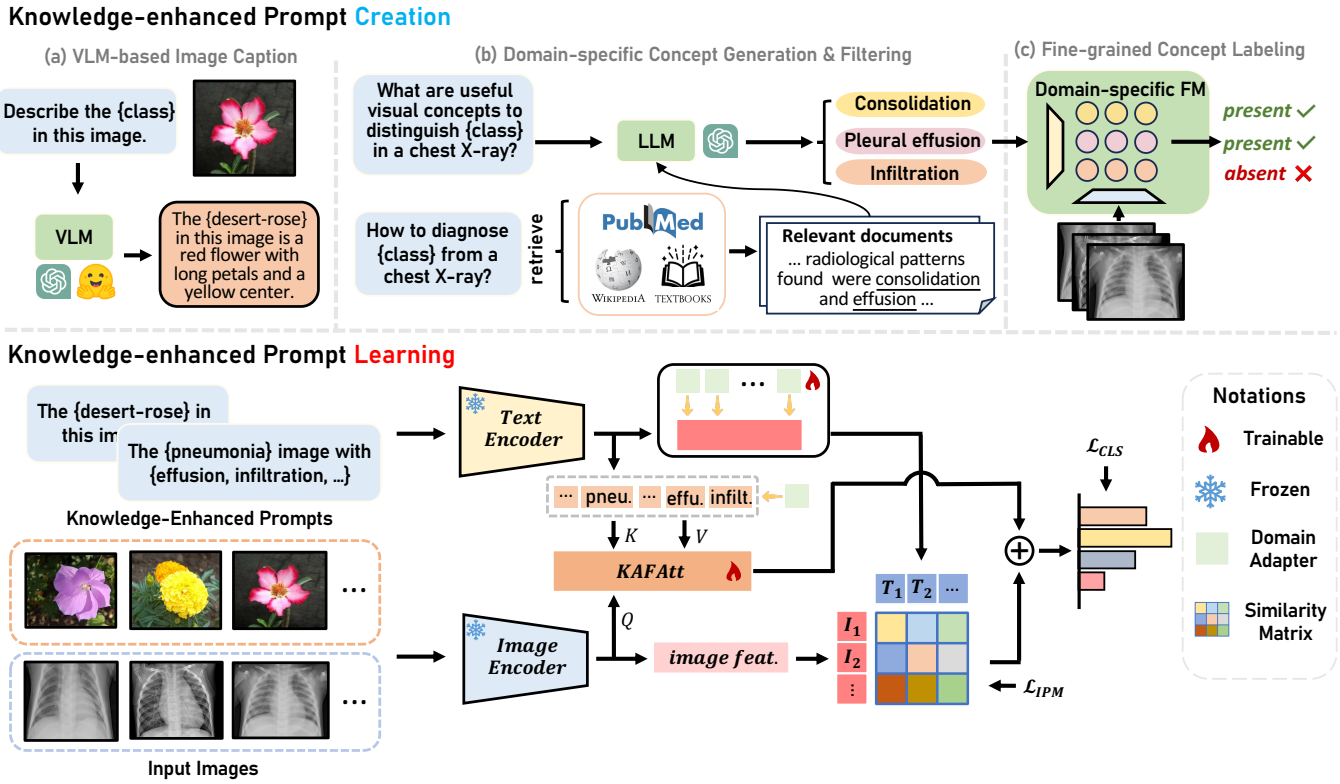

Figure 2: The overall pipeline of **KEEP**. The key insight of KEEP is improving both the performance and interpretability of the adaptation process for VLMs on various domains by introducing fine-grained knowledge elicited from domain-specific foundation models and RAG, highlighting the collaboration between FMs and XAI.

et al. 2021) leverages concept-based knowledge such as tail, beak, and head when classifying bird images and offers concept-based explanations. In healthcare, clinical knowledge is crucial when diagnosing diseases, e.g., Xiang et al. (2024) propose using ovarian–adnexal reports and routine clinical variables provided by radiologists to help predict ovarian cancers and improve model interpretability. In addition, retrieval-augmented generation (RAG) has emerged as an effective approach using large language models for knowledge-intensive tasks (Gao et al. 2023), which has been used in various domains (Xiong et al. 2024). In this work, we propose to incorporate RAG and domain-specific foundation models to provide more reliable image-wise knowledge for prompt learning in various domains, e.g., we use elicited clinical-concept-based knowledge for disease diagnosis of chest X-rays and brain MRI, etc., achieving both performance and explainability improvement.

3 Method

In this section, we first review the preliminaries of CLIP (Sect 3.1). Then we introduce our proposed **KEEP**, which mainly comprises two stages. The first stage is Knowledge-Enhanced Prompt Creation (Sect 3.2), where we utilize domain-specific FMs and retrieval-augmented generation to obtain fine-grained image-wise knowledge. The second stage is Knowledge-Enhanced Prompt Learning (Sect 3.3), which is the training pipeline of our explainable prompting framework aligning the images and the generated knowledge via a fine-grained attention mechanism and logit fusion.

3.1 Preliminaries

CLIP (Contrastive Language-Image Pre-training) is a representative foundational vision-language model that creates a shared embedding space through vision-language contrastive learning. CLIP consists of two encoders: a vision encoder $E_v(\cdot)$ that takes images as input and outputs the corresponding visual embeddings in the latent space, and a text encoder $E_t(\cdot)$ that maps the text input to the text embeddings. During inference, the input prompt of CLIP is *a photo of a* [*class name*], and the prediction probability is computed by the image-text similarity:

$$P(y=m|I) = \frac{\exp(\cos(E_v(I), E_t(P_m))/\tau)}{\sum_{j=1}^{M} \exp(\cos(E_v(I), E_t(P_j))/\tau)}, \quad (1)$$

where I is the image, m stands for the m-th class, P_m denotes the class prompt, M is the number of classes, $\cos(\cdot,\cdot)$ is the cosine similarity, and τ is a temperature parameter.

3.2 Knowledge-enhanced Prompt Creation

Knowledge is essential for bridging the gap between humans and AI models (Tocchetti et al. 2022). It empowers users to gain deeper insights into the underlying reasoning by enabling models to mimic the decision-making processes

Algorithm 1: Knowledg-Enhanced Prompt Creation

Input: A given image $\mathcal{I}$ and its class label $\mathcal{Y}_\mathcal{I}$, the domain-specific foundation model $\mathbf{DSFM}$.
Output: Knowledge-enhanced prompt $\mathcal{R}_\mathcal{I}$ for image $\mathcal{I}$.
$\mathcal{G}$: corpus (e.g., PubMed), $\mathcal{Q}$: queries, $\mathcal{C}$: set of candidate concepts, $\mathcal{C}_\mathcal{I}$: labeled concepts for image $\mathcal{I}$.
$\mathcal{P}$: set of positive and negative prompts for $\mathbf{DSFM}$.
$\mathcal{C}_1 \leftarrow \mathbf{LLM}(\mathcal{Q}(\mathcal{Y}_\mathcal{I}))$ `// LLM concepts`
$\mathcal{G}' \leftarrow \mathbf{Retrieve}(\mathcal{G}, \mathcal{Q}(\mathcal{Y}_\mathcal{I}))$
$\mathcal{C}_2 \leftarrow \mathbf{LLM}(\mathcal{G}')$ `// RAG concepts`
$\mathcal{C} \leftarrow \mathbf{Filtering}(\mathcal{C}_1, \mathcal{C}_2)$ `// concept filtering`
for *c in* $\mathcal{C}$ **do**
  $\mathcal{C}_\mathcal{I} \leftarrow \mathcal{C}_\mathcal{I} + \mathbf{DSFM}(\mathcal{I}, \mathcal{P}(c))$
$\mathcal{R}_\mathcal{I} \leftarrow \mathbf{Concat}(\mathcal{Y}_\mathcal{I}, \mathcal{C}_\mathcal{I})$

of human experts using domain knowledge. However, fine-grained annotating for specific data is very expensive and time-consuming, which needs human experts' efforts. To address the issues, we propose eliciting knowledge from expert foundation models, as illustrated in the upper part of Figure 2. Specifically, since the development of foundational VLMs and the image caption techniques for the natural image domain is mature (Zhou et al. 2020; Zhang et al. 2024b), we query the foundation models such as MiniGPT-4 and GPT-4 to generate the description of a given natural image. For example, we can query the foundation model with a prompt "Describe the [class name] in this image" and the model will generate corresponding descriptions.

However, existing natural field FMs have limited performance in other domains and it is hard for them to offer accurate information. To address this issue, we obtain knowledge by incorporating retrieval augmented generation and domain-specific FMs for specific domains. For instance, in the medical field, the fine-grained clinical concept-based prompt is adopted instead of directly using image captions, as illustrated in Algorithm 1. Clinical concepts are relevant attributes or symptoms of diseases, e.g., *consolidation* is a concept for pneumonia in chest X-rays. The clinical concepts of a given disease can be generated by prompting an LLM with queries such as "What are useful visual concepts to distinguish [disease name] in a {chest X-ray, brain MRI, etc.}?" Then RAG is adopted to improve the quality and reliability of the concepts. Given a corpus G covering various medical documents, e.g., PubMed (Canese et al. 2013), Wikipedia, and medical textbooks (Jin et al. 2021), we use prompts with specific disease names to retrieve relevant documents. The clinical concepts are extracted by an LLM and used to filter the originally generated concepts. To achieve an explainable framework that meticulously mimics the decision-making process of humans, we argue that class-level knowledge of previous methods is insufficient and coarse-grained, which cannot offer local explanations (Van der Velden et al. 2022). Medical experts diagnose diseases with domain knowledge case by case instead of limiting to generic knowledge. Inspired by this, we adopt domain-specific FMs (e.g., radiology domain) to give the predicted presence results of given clinical concepts for each image. Specifically, given the clinical candidate concepts $C = \{c_1, c_2, ..., c_{N_c}\}$ (N_c is the number of concepts) generated by LLM and RAG, an input image I, let $E_v(\cdot)$ and $E_t(\cdot)$ denote the vision and text encoder of the domain-specific FM, respectively, then the presence of a specific concept c_i is calculated by

$$Pre_{c_i} = \text{argmax}\{\text{sim}(E_v(I), E_t(N^{c_i})), \text{sim}(E_v(I), E_t(P^{c_i}))\}, \tag{2}$$

where $\text{sim}(\cdot)$ stands for the similarity, Pre_{c_i}=1 or Pre_{c_i}=0 represent concept c_i is present or absent in this image, P^{c_i} and N^{c_i} denote the positive (present) and negative (absent) prompt for concept c_i, respectively. The image-wise knowledge-enhanced prompts R are created by concatenating the present clinical concepts and category names of corresponding images, for example, a knowledge-enhanced prompt for a given dermoscopic image can be "a photo of melanoma, with irregular dots and globules, blue whitish veil". The reliability of the elicited knowledge is improved and demonstrated by RAG and knowledge intervention (Sect 4.3), where the retrieved results are consistent with the clinical fact (Kermany et al. 2018; Gezercan et al. 2016).

3.3 Knowledge-enhanced Prompt Learning

In the prompt learning process of our framework, image-wise knowledge is used as the input to the text encoder of the pre-trained vision-language model. The category of an object typically hinges on various visual concepts observable within specific, localized regions in an image. For example, in a chest X-ray of pneumonia, *consolidation* can be a distinguishable concept presented in some regions. Given that different concepts may correspond to distinct sub-regions of an image, we adopt a knowledge-aware fine-grained attention (KAFAtt). Specifically, the embeddings of the input images are linearly projected into the query matrix $Q \in (N, dim)$ while the key matrix and value matrix $K, V \in (N, dim)$ are the linear projections of the corresponding fine-grained text embeddings of the image-wise knowledge-enhanced prompts, where N and dim denote the number of samples and the dimension of embeddings, respectively. We can obtain the attention weight by normalizing the production of the query matrix and key matrix. The output of the KAFAtt module is the multiplication of the attention weights and the value matrix. A projection matrix is adopted to map the original embedding dimension to the number of classes M:

$$logit_{\text{IPA}} = Proj(softmax(\frac{QK^T}{\sqrt{dim}})V), \tag{3}$$

where $logit_{\text{IPA}}$ denotes the logit output by the image-prompt KAFAtt module, and $Proj(\cdot) : dim \rightarrow M$ stands for the linear projection layer. To explicitly preserve the prior knowledge and learn the generic knowledge from the specific domain, we propose using a domain adapter D instead of learning the original input prompts. As shown in Figure 2, the domain adapter is a learnable matrix added to the text embedding of the prompts, avoiding destroying the knowledge prior elicited from domain-specific foundation models, hence preserving the explainability of prompts. Then the prompt embedding is used for image-text matching through contrastive learning. A probability distribution over the class labels is given by:

Method	Pub.	Knowledge	Derm7pt (Dermoscopy)	CCBTM (Brain MRI)	Pneumonia (Chest X-ray)	Open-i (Chest X-ray)	Average
CLIP (Radford et al. 2021)	ICML'21	-	69.11	29.51	62.52	13.21	43.59
CoOp (Zhou et al. 2022b)	IJCV'22	✗	$73.94_{\pm 2.26}$	$78.02_{\pm 0.73}$	$85.06_{\pm 1.45}$	$71.72_{\pm 2.59}$	77.19
CoCoOp (Zhou et al. 2022a)	CVPR'22	✗	$75.37_{\pm 2.11}$	$84.32_{\pm 0.71}$	$85.96_{\pm 1.45}$	$70.80_{\pm 2.68}$	79.11
Tip-Adapter (Zhang et al. 2022)	ECCV'22	✗	$69.19_{\pm 2.22}$	$50.97_{\pm 0.88}$	$62.64_{\pm 1.91}$	$69.01_{\pm 2.66}$	62.95
Tip-Adapter-F (Zhang et al. 2022)	ECCV'22	✗	$68.93_{\pm 2.31}$	$72.26_{\pm 0.81}$	$79.84_{\pm 1.61}$	$69.21_{\pm 2.66}$	72.56
GraphAdapter (Li et al. 2024)	NeurIPS'23	✗	$75.78_{\pm 2.21}$	$82.35_{\pm 0.68}$	$85.53_{\pm 1.45}$	$73.26_{\pm 2.42}$	79.23
KgCoOp (Yao et al. 2023)	CVPR'23	✓	$74.09_{\pm 2.25}$	$67.06_{\pm 0.82}$	$82.05_{\pm 1.83}$	$70.58_{\pm 2.57}$	73.45
LASP (Bulat et al. 2023)	CVPR'23	✓	$76.91_{\pm 2.15}$	$\underline{90.89}_{\pm 0.51}$	$\underline{93.13}_{\pm 1.00}$	$76.23_{\pm 2.41}$	$\underline{84.29}$
TCP (Yao et al. 2024)	CVPR'24	✓	$\underline{76.96}_{\pm 2.13}$	$70.43_{\pm 0.79}$	$79.70_{\pm 1.60}$	$71.30_{\pm 2.55}$	74.60
BiomedCoOp (Koleilat et al. 2025)	CVPR'25	✓	$70.63_{\pm 2.29}$	$85.20_{\pm 0.62}$	$85.85_{\pm 1.40}$	$\underline{76.79}_{\pm 2.37}$	79.62
KEEP (Ours)	-	✓	$\mathbf{81.13}_{\pm 1.93}$	$\mathbf{95.12}_{\pm 0.36}$	$\mathbf{94.69}_{\pm 0.92}$	$\mathbf{77.42}_{\pm 2.36}$	**87.09**

Table 1: Quantitative comparison on disease diagnosis (classification) for medical image datasets with the state-of-the-art methods. In this paper, our medical image datasets include dermoscopic images, brain MRIs, and chest X-rays. The performance is reported as mean$_{\pm \mathrm{std}}$[%], where the std is derived from the 95% confidence interval. Our method is highlighted in grey. The best and the second-best results are shown in **bold** and underlined, respectively.

Method	Pub.	Knowledge	Caltech-101	Aircraft	Flowers	DTD	Average
CLIP (Radford et al. 2021)	ICML'21	-	92.94	24.60	71.34	44.44	58.33
CoOp (Zhou et al. 2022b)	IJCV'22	✗	$95.72_{\pm 0.40}$	$39.69_{\pm 0.82}$	$95.90_{\pm 0.39}$	$69.24_{\pm 1.11}$	75.14
CoCoOp (Zhou et al. 2022a)	CVPR'22	✗	$95.53_{\pm 0.42}$	$35.03_{\pm 0.83}$	$93.12_{\pm 0.52}$	$65.03_{\pm 1.21}$	72.18
Tip-Adapter (Zhang et al. 2022)	ECCV'22	✗	$94.83_{\pm 0.45}$	$40.07_{\pm 0.87}$	$94.54_{\pm 0.43}$	$66.08_{\pm 1.17}$	73.88
Tip-Adapter-F (Zhang et al. 2022)	ECCV'22	✗	$95.71_{\pm 0.42}$	$45.00_{\pm 0.83}$	$96.25_{\pm 0.39}$	$71.09_{\pm 1.11}$	77.01
GraphAdapter (Li et al. 2024)	NeurIPS'23	✗	$95.69_{\pm 0.42}$	$47.68_{\pm 0.84}$	$97.86_{\pm 0.30}$	$72.06_{\pm 1.07}$	78.32
KgCoOp (Yao et al. 2023)	CVPR'23	✓	$95.18_{\pm 0.43}$	$37.23_{\pm 0.78}$	$93.67_{\pm 0.57}$	$70.42_{\pm 1.07}$	74.12
LASP (Bulat et al. 2023)	CVPR'23	✓	$95.83_{\pm 0.40}$	$36.42_{\pm 0.82}$	$96.08_{\pm 0.38}$	$70.00_{\pm 1.09}$	74.58
TCP (Yao et al. 2024)	CVPR'24	✓	$95.57_{\pm 0.43}$	$43.75_{\pm 0.86}$	$97.72_{\pm 0.31}$	$72.88_{\pm 1.08}$	77.48
CoAPT (Lee et al. 2025)	KBS'25	✓	$\underline{96.20}_{\pm 0.39}$	$\underline{48.30}_{\pm 0.86}$	$\underline{97.99}_{\pm 0.28}$	$\underline{74.50}_{\pm 1.05}$	$\underline{79.25}$
KEEP (Ours)	-	✓	$\mathbf{96.98}_{\pm 0.34}$	$\mathbf{50.73}_{\pm 0.86}$	$\mathbf{98.44}_{\pm 0.24}$	$\mathbf{77.36}_{\pm 1.02}$	**80.88**

Table 2: Quantitative comparison on image classification for natural image datasets with the state-of-the-art methods. Natural image datasets here refer to images from normal RGB cameras, e.g., generic objects, aircrafts, flowers, and textures.

$$P(y=m|I) = \frac{\exp(\cos(E_v(I), F_m)/\tau)}{\sum_{j=1}^{M} \exp(\cos(E_v(I), F_j)/\tau)}, \quad (4)$$

where F_m is the prompt embeddings added with domain adapter D for class m, and τ is a temperature parameter. The final output logit of our framework is the fusion of the $logit_{\mathrm{IPA}}$ output by the image-prompt KAFAtt module and the image-prompt matching similarity $logit_{\mathrm{IPM}} = E_v(I)E_t(R)^T$. The overall objective $\mathcal{L}$ is the average of image-prompt contrastive loss and the cross-entropy classification loss $\mathcal{L}_{\mathrm{CLS}}$ which measures the discrepancy between the final fusion logits and the ground-truth labels y:

$$\mathcal{L} = \tfrac{1}{2}[\underbrace{-\sum_{j=1}^{M} \log P(y=j|I)}_{\mathcal{L}_{\mathrm{IPM}}} + \underbrace{CE(\beta \cdot logit_{\mathrm{IPA}} + (1-\beta) \cdot logit_{\mathrm{IPM}}), y)}_{\mathcal{L}_{\mathrm{CLS}}}], \quad (5)$$

where β is a logit-balanced hyperparameter, and $CE(\cdot)$ denotes the cross-entropy loss.

4 Experiments

4.1 Experimental Setups

Datasets. Our framework was evaluated on a comprehensive benchmark of 8 datasets spanning a diverse set of domains and image modalities, including (1) Dermoscopic images: *Derm7pt* (2018); (2) Brain MRI: *CCBTM* (2023); (3) Chest X-ray images: *Pneumonia* (2018), *Open-i* (2016); (4) Generic objects: *Caltech101* (2004); (5) Fine-grained images of flowers: *Oxford-Flowers102* (2008); (6) Fine-grained images of aircrafts: *FGVC-Aircraft* (2013) and (7) Images of textures: *DTD* (2014). To demonstrate that our method can be flexibly applied to datasets with and without knowledge annotations, the clinical concept annotations of *Derm7pt* were used to create the knowledge-enhanced prompts, while knowledge of domain-specific FMs was adopted for other datasets. We evaluate the test accuracy.

Implementation Details. Our framework adopted the pre-trained visual (ViT-B/16) and text encoder of CLIP. We adopted the SGD optimizer with a learning rate of 0.032. We used warm-up and cosine anneal as training strategies. All prompt learning methods implemented in this paper adopted random crop and random flip for data augmentation. Grid search was used to select hyperparameters, and β is set to 0.7. All comparison experiments were conducted on an RTX 4090 GPU. Image caption for natural images was based on MiniGPT-4 (Zhu et al. 2023). For medical RAG, we adopted the corpus organized by Xiong et al. (2024).

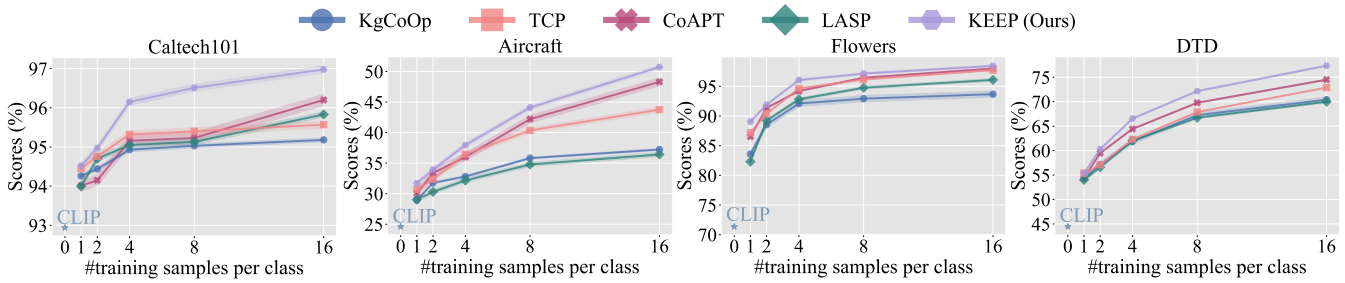

Figure 3: The few-shot results on natural image datasets. All methods are evaluated under 1, 2, 4, 8, and 16-shot settings.

Method	Derm7pt		CCBTM		Pneumonia		Open-i	
	10%	50%	10%	50%	10%	50%	10%	50%
KgCoOp	70.30	72.90	60.78	64.76	74.07	77.05	68.96	70.43
BiomedCoOp	69.84	69.35	84.05	84.37	88.05	87.09	74.61	75.89
TCP	70.97	75.34	69.02	69.08	79.03	79.42	70.41	71.27
LASP	72.37	76.31	82.69	91.29	87.86	91.52	71.81	75.53
KEEP (Ours)	**74.66**	**79.03**	**92.89**	**94.73**	**91.35**	**94.50**	**73.58**	**77.03**

Table 3: Experimental results on medical image datasets with different proportions of training data (10% and 50%).

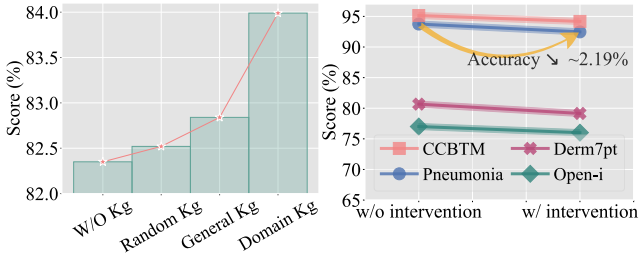

Figure 4: Illustration and results of knowledge intervention.

Method	Medical	Natural	Δ
KEEP	**87.09**	**80.88**	-
w/o $logit_{\text{IPA}}$	85.98	80.16	-0.9
w/o $\mathcal{L}_{\text{IPM}}$	84.69	79.61	-1.8
w/o $\mathcal{L}_{\text{CLS}}$	80.50	70.25	-8.6

Table 4: Ablation study of the fusion logits and losses. The average results of each field are reported.

4.2 Experimental Results

Results in various domains and image modalities. In Table 1, we report the disease diagnosis comparison results on four medical datasets of different modalities, including dermoscopic images, brain MRIs, and chest X-rays. The results on natural image datasets are shown in Table 2, including performance comparison for generic objects, fine-grained aircraft and flowers, and texture classification. Following previous methods, the results on natural image datasets are under the 16-shot setting. Note that we compare with the best version of CoAPT (based on DePT (2024c)) and BiomedCoOp (based on BiomedCLIP (2023a)) for natural and medical image data, respectively. CLIP baseline without any tuning is included in the first row. Our method outperforms other SOTA methods by a significant margin, gaining an average relative improvement of 3.3% on medical images and 2.1% on natural images over the second-best results. Note that CLIP shows an accuracy of 13.2% in *Open-i* and less than 45% overall accuracy in four medical datasets, exhibiting a significant domain gap, while our method improves performance through CLIP adaptation by an accuracy of 43.5% in the medical field, even consistently outperforming BiomedCoOp (which is based on BiomedCLIP with better performance than original CLIP) in all image modalities, demonstrating the effectiveness and robustness of our framework in handling tasks of diverse domains.

Data Efficiency. We conduct experiments to evaluate the effectiveness and data efficiency of **KEEP**. Specifically, for medical datasets (Table 3), we report the performance with different proportions of training data. It can be observed that **KEEP**, while showing the best results when using full data, does not exhibit significant declines when only 50% or 10% of the diagnosis labels are used on most datasets. For example, there is nearly no performance drop on *Pneumonia* dataset when the training data proportion drops from 100% to 50%. In addition, the performance of LASP drops from 91.3% to 82.7% on *CCBTM* dataset when the data proportion reduces from 50% to 10%, while our method exhibits much less performance gap (94.7% → 92.9%). For the natural images, few-shot learning is adopted to evaluate the efficiency (Figure 3). **KEEP** consistently outperforms other methods by a significant margin. For example, **KEEP** respectively gains 0.98%, 1.60%, 1.49%, 3.74%, 7.98% performance boost over TCP at 1, 2, 4, 8, and 16 shots on *Aircraft*, respectively. The consistent results in various domains indicate that our method encourages the model to learn the correspondences between images and fine-grained domain knowledge effectively, enabling the model to achieve promising adaptation performance and data efficiency.

Alabtion Study. We conduct ablation experiments for all the datasets on the effectiveness of the proposed image-prompt attention-based logit (i.e., $logit_{\text{IPA}}$, which is used to fuse with the original similarity logit), and the proposed losses (i.e., the image-prompt matching contrastive loss $\mathcal{L}_{\text{IPM}}$ and the cross-entropy $\mathcal{L}_{\text{CLS}}$ loss for fusion logits). As shown in Table 4, the overall performance drops significantly when removing the proposed components during the prompt learning process. Our method achieves the best overall perfor-

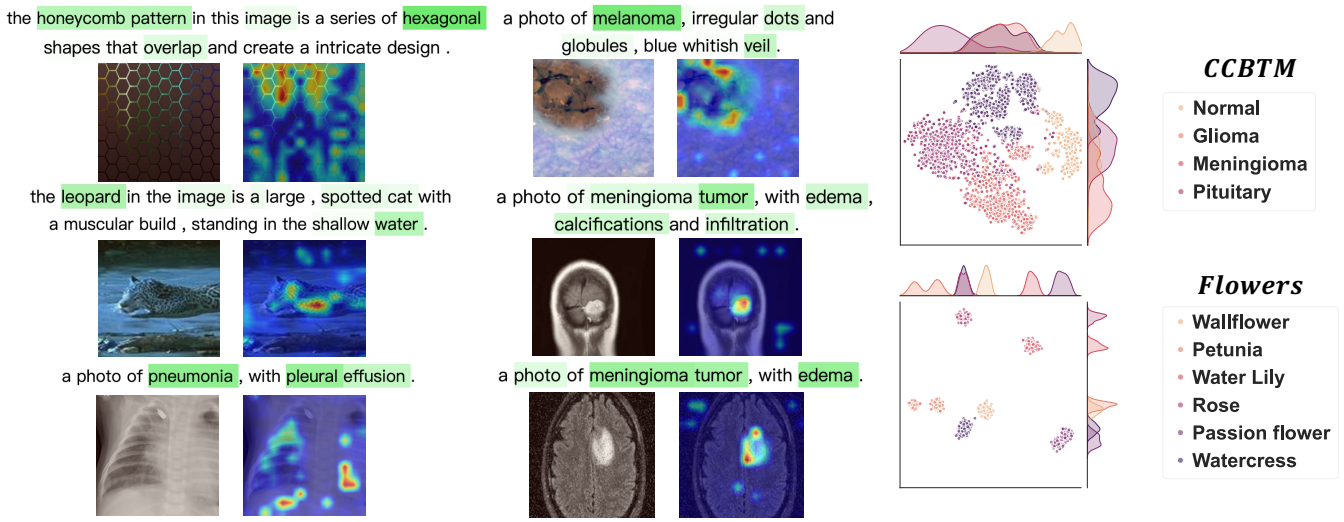

Figure 5: (a) Examples of image-prompt attention visualization. Darker (yellow) or lighter (blue) colors indicate higher or lower relevance scores between words of knowledge prompts and the image regions, respectively. (b) The t-SNE visualization of different domains, including *CCBTM* and *Flowers*. Six categories with the largest sample numbers are selected for *Flowers*.

mance across various domains with all components.

4.3 Analysis of Explainbility

In this section, we evaluate and analyze the explainability of our method. Drawing inspiration from prior research (Jin et al. 2023; Hsiao et al. 2021), we assess our framework using several essential metrics for XAI techniques, including *faithfulness*, *understandability*, and *plausibility*.

Faithfulness. *Faithfulness* is defined as the extent to which an explanation truthfully reflects the model's decision-making process, requiring the explanation to be highly faithful to the designed model mechanism (Lakkaraju et al. 2019; Rigotti et al. 2021). We evaluate *faithfulness* by intervening in the input knowledge (Kg)-enhanced prompts. Specifically, we use five kinds of prompt settings, including prompts without Kg, with random Kg (random tokens as prompts), with general Kg (prompts w/o domain-specific Kg), with our fine-grained domain-specific Kg, and the intervened Kg (the prompts semantics are modified, e.g., the descriptions of a normal instance may be replaced by that of an abnormal one or do the opposite like replacing "regular pigmentation" with "irregular pigmentation"). The left part of Figure 4 reports the overall performance of all eight datasets with different Kg settings, while the right part shows the Kg intervention results for medical datasets. These results show that not using Kg, using only random Kg, coarse-grained general Kg, or Kg after intervention as prompts may lead to performance degradation, demonstrating that the adopted domain knowledge faithfully explains the model's decisions and the knowledge reliability.

Understandability & Plausibility. *Understandability* requires explanations to be easily understandable to users without much technical knowledge (Jin et al. 2023; Johansson et al. 2004), while *plausibility* refers to how convincing the explanation appears (Hsiao et al. 2021). Our framework achieves *understandability* and *plausibility* by offering both visual and textual explanations, as shown in Figure 5(a). Specifically, we visualize the attention maps of images and their corresponding word importance of the Kg-enhanced prompts based on the predicted image-prompt matching logits. The results show that **KEEP** can accurately focus on meaningful and discriminative image regions and knowledge. For example, in the upper right case (dermoscopic image), "melanoma" is the correctly predicted disease label and is highlighted with the highest relevance score, while meaningful clinical concepts such as "dots" and "veils" are also highlighted. Figure 5(b) presents the t-SNE visualization of sample embeddings for our method. The well-clustered results highlight the strong distinguishing ability of our model in diverse domains. The provided explanations enhance human understanding of the model's decision-making process by clarifying the utilized knowledge and the focused areas. This can potentially assist domain experts in applying AI models to practical scenarios, e.g., helping medical professionals understand AI models for diagnosis.

5 Conclusion

In this paper, we propose **KEEP**, a knowledge-enhanced explainable prompting framework that leverages fine-grained domain-specific knowledge to enhance VLM adaptation in various domains, facilitating bridging the performance and interpretability gap of different domains. By incorporating image-wise knowledge elicited from domain-specific foundation models and meticulously learning the semantic correlations between images and knowledge-enhanced prompts, **KEEP** achieves promising performance and data efficiency, while improving interpretability by offering visual and textual explanations demonstrated by experiments and analysis conducted on diverse domains and image modalities, highlighting the collaboration between FMs and XAI.

Acknowledgments

This work was supported by the Pneumoconiosis Compensation Fund Board, HKSAR (Project No. PCFB22EG01), the Hong Kong Innovation and Technology Commission (Project No. GHP/006/22GD and ITCPD/17-9), HKUST (Project No. FS111), and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. T45-401/22-N).

References

Bie, Y.; Luo, L.; and Chen, H. 2024a. Mica: Towards explainable skin lesion diagnosis via multi-level image-concept alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 837–845.

Bie, Y.; Luo, L.; Chen, Z.; and Chen, H. 2024b. Xcoop: Explainable prompt learning for computer-aided diagnosis via concept-guided context optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 773–783. Springer.

Bulat, A.; and Tzimiropoulos, G. 2023. LASP: Text-to-Text Optimization for Language-Aware Soft Prompting of Vision & Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23232–23241.

Canese, K.; and Weis, S. 2013. PubMed: the bibliographic database. *The NCBI handbook*, 2(1).

Cao, Q.; Xu, Z.; Chen, Y.; Ma, C.; and Yang, X. 2024. Domain-controlled prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 936–944.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, 178–178. IEEE.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *arXiv preprint arXiv:2110.04544*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gezercan, Y.; Acik, V.; Çavuş, G.; Ökten, A. I.; Bilgin, E.; Millet, H.; and Olmaz, B. 2016. Six different extremely calcified lesions of the brain: brain stones. *Springerplus*, 5: 1–10.

Gu, J.; Han, Z.; Chen, S.; Beirami, A.; He, B.; Zhang, G.; Liao, R.; Qin, Y.; Tresp, V.; and Torr, P. 2023. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*.

Hashemi, M. H. 2023. Crystal clean: Brain tumors MRI dataset. *Kaggle (accessed 09 May 2023)*.

Hou, J.; Liu, S.; Bie, Y.; Wang, H.; Tan, A.; Luo, L.; and Chen, H. 2024. Self-explainable ai for medical image analysis: A survey and new outlooks. *arXiv preprint arXiv:2410.02331*.

Hsiao, J. H.-w.; Ngai, H. H. T.; Qiu, L.; Yang, Y.; and Cao, C. C. 2021. Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI). *arXiv preprint arXiv:2108.01737*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

Jin, W.; Li, X.; Fatehi, M.; and Hamarneh, G. 2023. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Medical image analysis*, 84: 102684.

Johansson, U.; König, R.; and Niklasson, L. 2004. The truth is in there-rule extraction from opaque models using genetic programming. In *FLAIRS*, 658–663.

Kan, B.; Wang, T.; Lu, W.; Zhen, X.; Guan, W.; and Zheng, F. 2023. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15670–15680.

Kawahara, J.; Daneshvar, S.; Argenziano, G.; and Hamarneh, G. 2018. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2): 538–546.

Kermany, D. S.; Goldbaum, M.; Cai, W.; Valentim, C. C.; Liang, H.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5): 1122–1131.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19113–19122.

Koleilat, T.; Asgariandehkordi, H.; Rivaz, H.; and Xiao, Y. 2025. Biomedcoop: Learning to prompt for biomedical vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14766–14776.

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Leskovec, J. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 131–138.

Lee, G.; An, S.; Baik, S.; and Lee, S. 2025. CoAPT: Context Attribute words for prompt tuning. *Knowledge-Based Systems*, 113653.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, X.; Lian, D.; Lu, Z.; Bai, J.; Chen, Z.; and Wang, X. 2024. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Nie, Y.; He, S.; Bie, Y.; Wang, Y.; Chen, Z.; Yang, S.; Cai, Z.; Wang, H.; Wang, X.; Luo, L.; et al. 2025. An Explainable Biomedical Foundation Model via Large-Scale Concept-Enhanced Vision-Language Pre-training. *arXiv preprint arXiv:2501.15579*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. IEEE.

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

Rigotti, M.; Miksovic, C.; Giurgiu, I.; Gschwind, T.; and Scotton, P. 2021. Attention-based interpretability with concept transformers. In *International conference on learning representations*.

Tan, H.; Li, J.; Zhou, Y.; Wan, J.; Lei, Z.; and Zhang, X. 2024. Compound text-guided prompt tuning via image-adaptive cues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5061–5069.

Tocchetti, A.; and Brambilla, M. 2022. The role of human knowledge in explainable AI. *Data*, 7(7): 93.

Van der Velden, B. H.; Kuijf, H. J.; Gilhuijs, K. G.; and Viergever, M. A. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79: 102470.

Xiang, H.; Xiao, Y.; Li, F.; Li, C.; Liu, L.; Deng, T.; Yan, C.; Zhou, F.; Wang, X.; Ou, J.; et al. 2024. Development and validation of an interpretable model integrating multimodal information for improving ovarian cancer diagnosis. *Nature Communications*, 15(1): 2681.

Xiong, G.; Jin, Q.; Lu, Z.; and Zhang, A. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Yao, H.; Zhang, R.; and Xu, C. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6757–6767.

Yao, H.; Zhang, R.; and Xu, C. 2024. TCP: Textual-based Class-aware Prompt tuning for Visual-Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23438–23448.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. arXiv 2022. *arXiv preprint arXiv:2205.01917*.

Yu, T.; Lu, Z.; Jin, X.; Chen, Z.; and Wang, X. 2023. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10899–10909.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024a. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024b. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; Wu, S.; Gao, L.; Shen, H. T.; and Song, J. 2024c. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12924–12933.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, 493–510. Springer.

Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. 2023a. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Zhang, X.; Wu, C.; Zhang, Y.; Xie, W.; and Wang, Y. 2023b. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1): 4542.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; and Gao, J. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13041–13049.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.