KEEP: TOWARDS A KNOWLEDGE-ENHANCED EXPLAINABLE PROMPTING FRAMEWORK FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale vision-language models (VLMs) embedded with expansive representations and visual concepts have showcased significant potential in the computer vision community. Efficiently adapting VLMs such as CLIP, to downstream tasks 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 tasks that require domain-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 across various domains, facilitating bridging the gap between the general domain and other specific domains. We present to our best knowledge the first work to incorporate retrieval augmented generation and domain-specific foundation models to 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, demonstrate that our method simultaneously achieves superior performance and interpretability, shedding light on the effectiveness of the collaboration between foundation models and XAI. The code will be made publically available.

1 INTRODUCTION

Recent studies in large-scale vision-language pre-trained models (VLMs), such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022), ALIGN (Jia et al., 2021), Flamingo (Alayrac et al., 2022) 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 tasks like image classification, segmentation, and image-text retrieval (Lüddecke & Ecker, 2022; Fang et al., 2021). However, the massive sizes and high training costs have prompted researchers to explore efficient methods for adapting the pre-trained VLMs to downstream tasks.

Recently, prompt learning (Zhou et al., 2022a;b), which is introduced from the field of natural language processing, has emerged as one of the representative approaches for efficiently adapting foundation models to downstream tasks like image classification. These methods focus on learning the prompts instead of training all the parameters of the models, 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), which demonstrates relatively low classification accuracy when handling fine-grained tasks. Some studies tend to alleviate this issue by introducing knowledge into prompt learning (Yao et al., 2023; Bulat & Tzimiropoulos, 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-specific knowledge. The coarse-grained and insufficient information em-

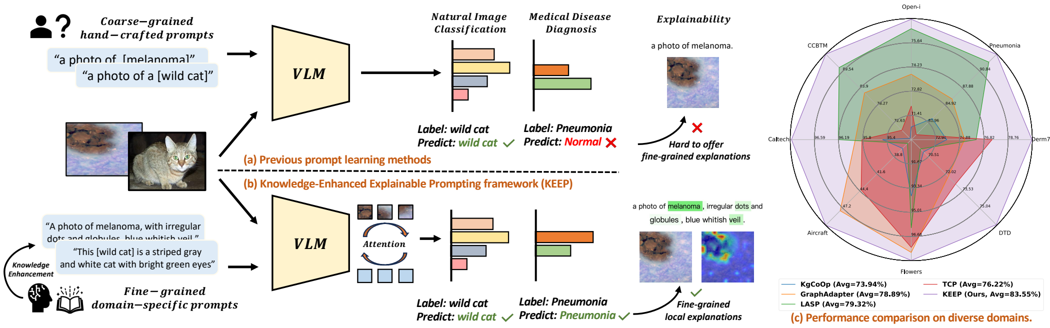


Figure 1: Illustration of Knowledge-Enhanced Explainable Prompting framework (**KEEP**) for various domains: (a) Previous works adopt only coarse-grained general prompts and usually perform well in limited domains. (b) **KEEP** utilizes domain knowledge-enhanced prompts to facilitate bridging the gap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

bedded in these models leads to unsatisfactory interpretability and cannot meet the trustworthiness requirements of XAI, especially in high-stakes scenarios such as healthcare (Hulsen, 2023).

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

We summarize our main contributions as follows: (i) We propose a knowledge-enhanced explainable prompting framework that leverages fine-grained domain-specific knowledge to enhance the VLM adaption. An image-prompt attention module is further proposed 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 including different modalities from medical and natural fields. (iii) Extensive experiments and explainability analyses show that our method concurrently achieves promising performance and interpretability. To the best of our knowledge, we are the first to explore using image-wise fine-grained knowledge elicited from domain-specific foundation models and RAG for prompt learning in various fields including medical and natural domains, 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 datasets, which ingest images and their respective textual descriptions as inputs and learn to associate the knowledge from the two modalities. According to the objectives, VLMs can be categorized as models with contrastive-only objectives (Radford et al., 2021; Li et al., 2021; Jia et al., 2021), generative objectives (Li et al., 2022; 2023; Bao et al., 2021), and alignment objectives (Singh et al., 2022; Dou et al., 2022). These models are usually built and extended from the following aspects: adopting stronger visual encoders (typically ResNet (He et al., 2016) or ViT (Dosovitskiy et al., 2020)) and textual encoders (typically transformer-based models (Vaswani, 2017)), e.g., BLIP2 (Li et al., 2023), training on larger datasets with image-text pairs (Schuhmann et al., 2022; Jia et al., 2021), and further fusing the visual and textual knowledge (Singh et al., 2022). Among existing vision-language models, CLIP (Radford et al., 2021) is one of the most representative and commonly used frameworks aligning the feature spaces of vision and text encoder via contrastive learning based on around 400 million image-text pairs.

Recently, the application of large-scale pre-trained vision-language models in other domains such as healthcare attracts increasing attention. These domain-specific foundation models aim to introduce the vision-language learning approach to medical vision and text understanding, facilitating building potential models for disease diagnosis (Tiu et al., 2022; Zhang et al., 2023b), medical VQA (Thawkar et al., 2023; Moor et al., 2023), and report generation (Pellegrini et al., 2023), etc. For example, MedCLIP (Wang et al., 2022) adopts contrastive learning for diagnosing chest X-ray images. 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 adaption methods from the field of natural language processing (Liu et al., 2023b). Prompt learning, especially soft prompt learning, aims to improve the adaption ability of VLMs 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 (i.e., *a photo of a* [*class name*]). For instance, CoOp (Context Optimization) (Zhou et al., 2022b) proposes to replace the fixed hand-crafted prompts with soft/learnable prompts and optimize the textual tokens. CoCoOp (Conditional Context Optimization) (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 (e.g., *a photo of a* [*class name*]) show low performance when handling fine-grained tasks. Some recent studies, e.g., KgCoOp (Yao et al., 2023), LASP (Bulat & Tzimiropoulos, 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. Therefore, we propose leveraging image-wise domain-specific knowledge to enhance the adaptation process, while improving model interpretability by providing prompt-based explanations.

2.3 Knowledge-based XAI

Bridging the understandability gap between humans and black-box AI models necessitates developing 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, which is derived from human knowledge in various fields, plays an important role in improving the model performance and explainability (Tocchetti & Brambilla, 2022). For example, Concept Transformer (Rigotti 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, data system scores, 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; Lewis et al., 2020), which has been used in various domains (Xiong et al., 2024; Liu et al., 2023a). We present to our best knowledge the first work 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 Approach

In this section, we first review the preliminaries (3.1) of CLIP (Radford et al., 2021). Then we introduce our proposed framework **KEEP**, which mainly comprises two stages. The first stage is Knowledge-Enhanced Prompt Creation (3.2), where we utilize domain-specific foundation models and retrieval-augmented generation to obtain fine-grained image-wise knowledge. The second stage is Knowledge-Enhanced Prompt Learning (3.3), which is the training pipeline of our explainable prompting framework aligning the images and the generated knowledge via an attention mechanism.

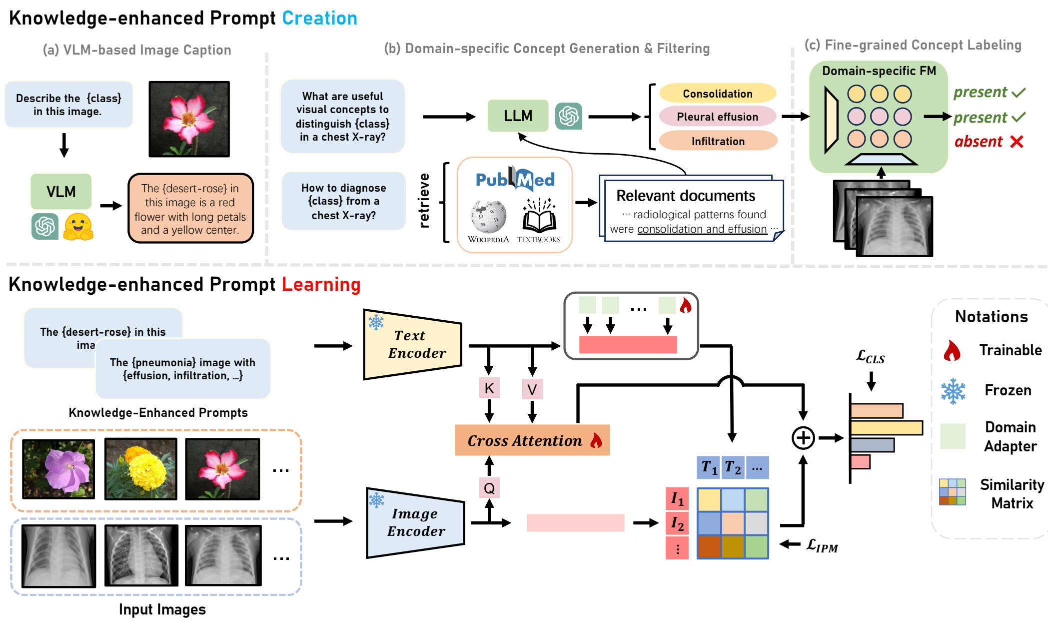


Figure 2: The overall pipeline of **KEEP**. The proposed framework comprises two stages: Knowledge-enhanced Prompt Creation and Knowledge-enhanced Prompt Learning. The key insight of KEEP is improving both the performance and interpretability of the adaption process for VLMs on various domains by introducing fine-grained knowledge elicited from domain-specific foundation models and RAG, highlighting the collaboration between FMs and XAI.

3.1 Preliminaries

CLIP (Contrastive Language-Image Pre-training (Radford et al., 2021)) 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:

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

where I represents the input image, m stands for the m-th class, P_m denotes the prompt for class m, 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 & Brambilla, 2022). It empowers users to gain deeper insights into the underlying reasoning by enabling models to mimic the decision-making processes of human experts using domain-specific knowledge. However, fine-grained annotating for specific data is very expensive and time-consuming, which needs human experts' efforts. To introduce domain-specific knowledge into the prompt learning process and alleviate the challenge of the high cost of knowledge annotations, we propose eliciting knowledge from expert foundation models, as illustrated in the upper part of Figure 2. Specifically, since the development of foundational vision-language models and the image caption techniques for the natural image domain is mature (Zhou et al., 2020; Zhang et al., 2024b;a), we query the foundation models such as MiniGPT-4 (Zhu et al., 2023) and GPT-4 (Achiam et al., 2023) 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 domain foundation models 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 foundation models for specific domains. For instance, in the medical domain, 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., pleural effusion is a clinical concept for pneumonia in chest X-rays. The clinical concepts of a given disease can be generated by prompting a large language model (LLM) with queries such as "What are useful visual concepts to distinguish [disease name] in a $\{$chest X-ray, dermoscopic image, 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 & Weis, 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 (Bulat & Tzimiropoulos, 2023; Yao et al., 2024) 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 foundation models (e.g., the 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 and negative 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 (Section 4.3). More details are in the appendix Section B.

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: The 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 $\mathbf{DSFM}$, see Section B.2 for details.
$\mathcal{C}_1 \leftarrow \mathbf{LLM}(\mathcal{Q}(\mathcal{Y}_\mathcal{I}))$ `// candidate concepts generated from LLM`
$\mathcal{G}' \leftarrow \mathbf{Retrieve}(\mathcal{G}, \mathcal{Q}(\mathcal{Y}_\mathcal{I}))$ `// retrieve relevant documents`
$\mathcal{C}_2 \leftarrow \mathbf{LLM}(\mathcal{G}')$ `// candidate concepts generated from RAG`
$\mathcal{C} \leftarrow \mathbf{Filtering}(\mathcal{C}_1, \mathcal{C}_2)$ `// filter the candidate concepts`
for c *in* $\mathcal{C}$ **do**
    $\mathcal{C}_\mathcal{I} \leftarrow \mathcal{C}_\mathcal{I} + \mathbf{DSFM}(\mathcal{I}, \mathcal{P}(c))$ `// image-wise concept labeling`
end
$\mathcal{R}_\mathcal{I} \leftarrow \mathbf{Concat}(\mathcal{Y}_\mathcal{I}, \mathcal{C}_\mathcal{I})$ `// the knowledge-enhanced prompt`

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 adopted an image-prompt attention module. 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 text embeddings, 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 image-prompt attention 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 query-key-value image-prompt attention 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 training the original input prompts. The domain adapter is a learnable matrix added to the text embeddings of the original class-level 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 :

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

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 $logit_{\text{IPA}}$ output by the image-prompt attention module and the image-prompt matching similarity $logit_{\text{IPM}} = E_v(I)E_t(R)^T$. The overall objective $\mathcal{L}$ is the average of image-prompt contrastive loss and the cross-entropy classification loss $\mathcal{L}_{\text{CLS}}$ which measures the discrepancy between the final fusion logits and the ground-truth labels y:

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

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

4 Experients

4.1 Experimental Setups

Datasets. Our framework was evaluated on a comprehensive benchmark of 8 datasets spanning a diverse set of domains, including (1) Dermoscopic images: *Derm7pt* (Kawahara et al., 2018); (2) Chest X-ray images: *Pneumonia* (Kermany et al., 2018), *Open-i* (Demner-Fushman et al., 2016); (3) Brain magnetic resonance imaging (MRI): *CCBTM* (Hashemi, 2023); (4) Generic objects: *Caltech101* (Fei-Fei et al., 2004); (5) Fine-grained images of flowers: *Oxford-Flowers102* (Nilsback & Zisserman, 2008); (6) Fine-grained images of aircraft: *FGVC-Aircraft* (Maji et al., 2013) and (7) Images of textures: *DTD* (Cimpoi et al., 2014). It should be noticed that 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 foundation models was adopted for other datasets. The accuracy of test sets was used for evaluation. Dataset and concept details are in the appendix Section A.

Baselines. We compared our model with classic and state-of-the-art adapter-based and prompt learning methods, including CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), Tip-Adapter (Zhang et al., 2022), Tip-Adapter-F (Zhang et al., 2022), KgCoOp (Yao et al., 2023), LASP (Bulat & Tzimiropoulos, 2023), GraphAdapter (Li et al., 2024), and TCP (Yao et al., 2024).

Implementation Details. Our framework adopted the pre-trained visual (ViT-B/16) and text encoder of CLIP (Radford et al., 2021). 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

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, chest X-rays, and brain MRIs. The performance is reported as mean_{std} of three random runs [%]. Our method is highlighted in light cyan. The best and the second-best results are shown in **bold** and underlined, respectively.

METHOD	Derm7pt	Pneumonia	Open-i	CCBTM	Average
CLIP	69.11	62.52	13.21	29.51	43.59
CoOp	$75.19_{\pm 0.36}$	$85.88_{\pm 0.56}$	$71.93_{\pm 0.71}$	$79.31_{\pm 0.84}$	$78.08_{\pm 0.62}$
CoCoOp	$77.04_{\pm 0.72}$	$86.06_{\pm 0.78}$	$70.63_{\pm 0.54}$	$84.67_{\pm 0.32}$	$79.60_{\pm 0.59}$
Tip-Adapter	$69.11_{\pm 0.00}$	$62.50_{\pm 0.00}$	$68.98_{\pm 0.00}$	$50.78_{\pm 0.08}$	$62.84_{\pm 0.02}$
Tip-Adapter-F	$69.11_{\pm 0.00}$	$81.25_{\pm 0.91}$	$69.31_{\pm 0.00}$	$73.88_{\pm 0.45}$	$73.39_{\pm 0.34}$
KgCoOp	$73.84_{\pm 1.37}$	$82.64_{\pm 0.30}$	$70.74_{\pm 1.21}$	$67.41_{\pm 0.38}$	$73.66_{\pm 0.82}$
GraphAdapter	$75.27_{\pm 1.86}$	$86.05_{\pm 0.13}$	$73.81_{\pm 0.41}$	$82.38_{\pm 0.11}$	$79.38_{\pm 0.63}$
LASP	$76.20_{\pm 1.56}$	$\underline{92.41_{\pm 0.08}}$	$\underline{76.46_{\pm 0.68}}$	$\underline{90.73_{\pm 0.33}}$	$\underline{83.95_{\pm 0.66}}$
TCP	$\underline{77.47_{\pm 0.20}}$	$79.86_{\pm 0.40}$	$71.95_{\pm 0.47}$	$70.09_{\pm 0.18}$	$74.84_{\pm 0.31}$
KEEP (Ours)	$\mathbf{80.67}_{\pm 0.31}$	$\mathbf{93.75}_{\pm 0.26}$	$\mathbf{77.01}_{\pm 0.31}$	$\mathbf{95.14}_{\pm 0.11}$	$\mathbf{86.64}_{\pm 0.24}$

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, where we include domains of generic objects, aircraft, flowers, and textures in this paper.

METHOD	Caltech-101	Aircraft	Flowers	DTD	Average
CLIP	92.94	24.60	71.34	44.44	58.33
CoOp	$95.87_{\pm 0.10}$	$39.05_{\pm 0.85}$	$95.75_{\pm 0.31}$	$68.93_{\pm 0.48}$	$74.90_{\pm 0.44}$
CoCoOp	$95.22_{\pm 0.28}$	$36.03_{\pm 0.21}$	$93.84_{\pm 0.21}$	$65.60_{\pm 0.42}$	$72.67_{\pm 0.28}$
Tip-Adapter	$94.74_{\pm 0.20}$	$39.24_{\pm 0.43}$	$93.90_{\pm 0.31}$	$65.76_{\pm 0.33}$	$73.41_{\pm 0.32}$
Tip-Adapter-F	$95.74_{\pm 0.03}$	$45.04_{\pm 0.77}$	$96.73_{\pm 0.20}$	$72.22_{\pm 0.35}$	$77.43_{\pm 0.34}$
KgCoOp	$95.47_{\pm 0.05}$	$37.43_{\pm 0.16}$	$93.88_{\pm 0.52}$	$70.08_{\pm 0.36}$	$74.22_{\pm 0.27}$
GraphAdapter	$95.92_{\pm 0.14}$	$\underline{47.63_{\pm 0.63}}$	$\underline{97.78_{\pm 0.13}}$	$72.26_{\pm 0.15}$	$\underline{78.40_{\pm 0.26}}$
LASP	$\underline{96.20_{\pm 0.07}}$	$36.61_{\pm 0.33}$	$96.07_{\pm 0.23}$	$69.82_{\pm 0.15}$	$74.68_{\pm 0.20}$
TCP	$95.81_{\pm 0.09}$	$44.20_{\pm 0.40}$	$97.43_{\pm 0.07}$	$\underline{72.91_{\pm 0.31}}$	$77.59_{\pm 0.22}$
KEEP (Ours)	$\mathbf{96.97}_{\pm 0.09}$	$\mathbf{49.99}_{\pm 0.35}$	$\mathbf{98.33}_{\pm 0.17}$	$\mathbf{76.50}_{\pm 0.78}$	$\mathbf{80.45}_{\pm 0.35}$

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 retrieval-augmented generation, we adopted the corpus organized by MEDRAG (Xiong et al., 2024), e.g., PubMed (Canese & Weis, 2013) and medical textbooks (Jin et al., 2021), for medical datasets. We used PMC-LLaMA 13B (Wu et al., 2024) as the LLM and MedCPT (Jin et al., 2023a) as the retriever for RAG. For domain-specific foundation models in the medical domain, we adopted KAD (Zhang et al., 2023b) and BiomedCLIP (Zhang et al., 2023a) to generate domain knowledge. More details can be found in the appendix Section B.2.

4.2 Experimental Results.

In order to comprehensively demonstrate the competitive performance of our method in both clinical disease diagnosis and natural image classification, comparison experiments with other state-of-the-art methods and ablation experiments on eight datasets of diverse domains are conducted.

Table 3: Experimental results on medical image datasets with different proportions of training data, including 10%, 50%, and 100%. Our method is highlighted in **bold**.

METHOD	Derm7pt			Pneumonia			Open-i			CCBTM		
	10%	50%	100%	10%	50%	100%	10%	50%	100%	10%	50%	100%
KgCoOp	70.89	72.91	73.84	74.36	76.60	82.64	68.98	70.63	70.74	60.90	64.77	67.41
TCP	71.65	75.69	77.47	79.49	79.33	79.89	70.29	71.95	71.95	69.26	69.47	70.09
GraphAdapter	69.11	69.37	75.27	65.54	85.73	86.05	68.98	71.28	73.81	74.70	81.62	82.38
LASP	72.15	75.94	76.20	87.50	91.34	92.41	71.62	74.59	76.46	82.72	91.54	90.73
KEEP (Ours)	**73.42**	**77.72**	**80.67**	**90.86**	**93.75**	**93.75**	**71.62**	**76.90**	**77.01**	**92.01**	**94.95**	**95.14**

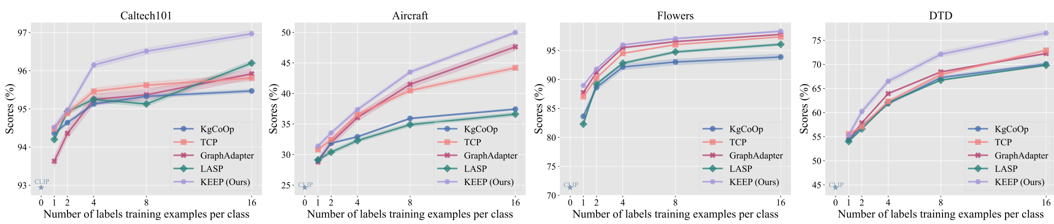


Figure 3: The few-shot learning results on four natural image datasets. All methods are evaluated under 1, 2, 4, 8, and 16-shot settings.

Results of Medical Image Diagnosis & Natural Image Classification. In Table 1, we report the disease diagnosis comparison results of our method on four medical datasets of different modalities, including dermoscopy images, chest X-ray images, and brain MRIs. The image classification 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 (Zhou et al., 2022b; Li et al., 2024), the results on natural image datasets are under the 16-shot setting. CLIP baseline (Radford et al., 2021) without any tuning is included at the first row of the two tables. Our method outperforms other state-of-the-art prompt learning methods by a significant margin, achieving an average relative improvement of approximately 3.2% on four medical datasets and 2.6% on four natural image datasets compared to the second-best results, which demonstrates the effectiveness and robustness of our framework in handling tasks across diverse domains.

Data Efficiency. To demonstrate the effectiveness and efficiency of our proposed framework, we conduct experiments to evaluate the data efficiency. Specifically, for the four medical image datasets, we report the performance with different proportions of training data, including 10%, 50%, and 100%, as shown in Table 3. We compare our method **KEEP** with state-of-the-art methods and it can be observed that the diagnosis performance of our method, 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 medical image 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 diagnosis results of LASP (Bulat & Tzimiropoulos, 2023) drop from 91.5% to 82.7% on *CCBTM* (Hashemi, 2023) dataset when the training data proportion reduces from 50% to 10%, while our method exhibits much less performance gap (i.e., from 94.9% to 92.0%). For the four natural image datasets, few-shot learning is adopted to evaluate the efficiency, including 1, 2, 4, 8, and 16 shots, as shown in Figure 3. Our method can consistently outperform other methods by a significant margin in most settings. For example, **KEEP** respectively gains 2.58%, 1.56%, 1.29%, 2.01%, 2.36% performance boost over GraphAdapter (Li et al., 2024) and outperforms TCP (Yao et al., 2024) by 0.57%, 1.11%, 0.81%, 3.03%, 5.79% at 1, 2, 4, 8, and 16 shots on *Aircraft* dataset, 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, thus facilitating the adaptation and enabling the model to achieve promising performance and data efficiency.

Alabtion Study. We conduct ablation experiments for all eight datasets on the effectiveness of the proposed image-prompt attention-based logit (i.e., $logit_{\mathrm{IPA}}$, which is used to fuse with the original

Table 4: Ablation study of the fusion logits and losses. MED. and NAT. represents the medical field and natural field, respectively. The average results of four datasets in each corresponding field are reported.

METHOD	MED.	NAT.	Δ
KEEP	**86.64**	**80.45**	-
w/o $logit_{\mathrm{IPA}}$	86.02	79.30	-0.9
w/o $\mathcal{L}_{\mathrm{IPM}}$	84.50	80.14	-1.2
w/o $\mathcal{L}_{\mathrm{CLS}}$	80.19	70.09	-8.4

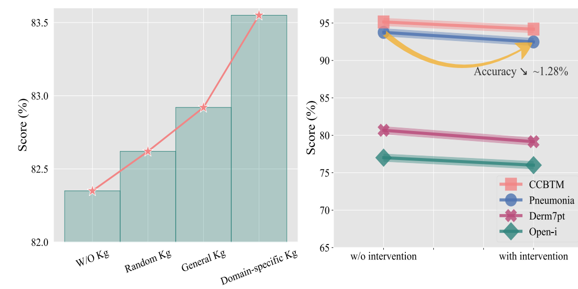


Figure 4: Illustration of our framework's faithfulness using knowledge intervention.

similarity logit), and the proposed losses (i.e., the image-prompt matching contrastive loss $\mathcal{L}_{\mathrm{IPM}}$ and the cross-entropy $\mathcal{L}_{\mathrm{CLS}}$ 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 performance across various domains with all designed components. More ablation results are in the appendix Section C.

4.3 Analysis of Explainability

In this section, we evaluate and analyze the explainability of our method. Drawing inspiration from prior research (Jin et al., 2023b; Hsiao et al., 2021; Guidotti et al., 2018; Johansson et al., 2004; Rigotti 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; Jin et al., 2023b). In this paper, we evaluate *faithfulness* by intervening the input knowledge-enhanced prompts. Specifically, we use five kinds of prompt settings, including prompts without knowledge, with random knowledge (i.e., random tokens as prompts), with general knowledge (i.e., prompts without domain-specific knowl-

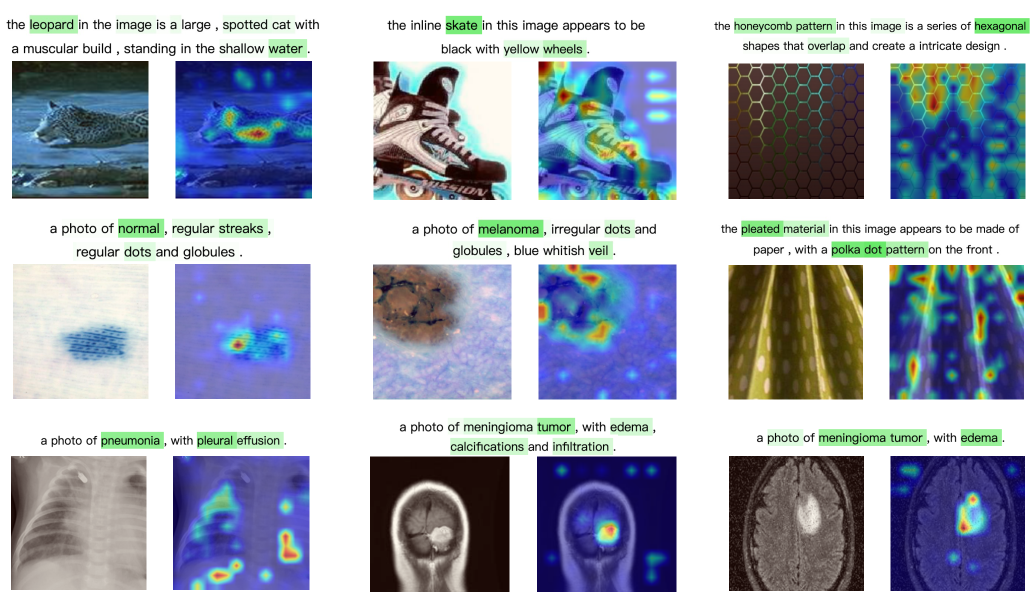


Figure 5: Examples of image-prompt attention visualization in various domains. Darker (yellow) or lighter (blue) colors indicate higher or lower relevance scores, respectively.

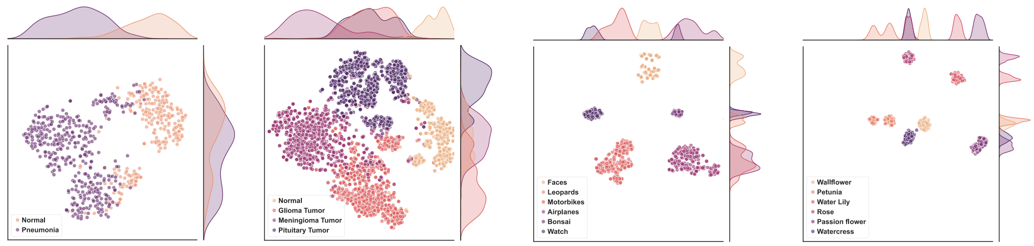


Figure 6: The t-SNE visualization results of different domains, including *Pneumonia*, *CCBTM*, *Caltech101*, and *Oxford-Flowers* datasets (from left to right). The six categories with the largest number of samples are selected for *Caltech101* and *Oxford-Flowers* datasets.

edge), with our fine-grained domain-specific knowledge and the intervened knowledge (intervened knowledge means that the semantics of the prompts are modified, e.g., the descriptions of a normal instance may be replaced by the descriptions 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 knowledge settings, while the right part shows the knowledge intervention results for medical image datasets. These results show that not using knowledge, using only random knowledge, coarse-grained general knowledge, or knowledge after intervention as prompts may lead to performance degradation, which demonstrates 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., 2023b; Johansson et al., 2004), while *plausibility* refers to how convincing the explanation appears (Hsiao et al., 2021; Jin et al., 2023b). Our framework achieves *understandability* and *plausibility* by offering both visual and textual explanations, as shown in Figure 5. Specifically, we visualize the attention maps of images and their corresponding word importance of the knowledge-enhanced prompts based on the predicted image-prompt matching logits and back-propagated gradients during training. The results show that our method can accurately focus on meaningful and discriminative image regions and knowledge. For example, in the middle case of Figure 5 (i.e., the case of dermoscopic image), “melanoma” is the correctly predicted disease label and is highlighted with the highest relevance score. Additionally, meaningful clinical concepts such as “dots”, ”globules” and “veils” are also highlighted by our method. Figure 6 presents the t-SNE visualization of sample embeddings for our method in various datasets, where different colors represent different categories and the embeddings cluster well. These results highlight the strong distinguishing ability of our model in diverse domains, benefiting from the semantic correlations between images and fine-grained domain knowledge. The explanations provided by our framework enhance human understanding of the model’s decision-making process by clarifying the utilized knowledge and the specific areas of focus. This can potentially assist domain experts in applying AI models to practical scenarios, such as helping medical professionals understand AI models for disease diagnosis.

5 Conclusion

In this paper, we propose **KEEP**, a knowledge-enhanced explainable prompting framework that leverages fine-grained domain-specific knowledge to enhance the adaptation process for VLMs in various domains, facilitating bridging the gap between the general domain and other specific domains. By incorporating domain knowledge elicited from domain-specific foundation models and meticulously learning the semantic correlations between images and knowledge-enhanced prompts based on the attention mechanism, our framework achieves promising performance and data efficiency, while improving interpretability by offering visual and textual explanations. The reliability of the elicited knowledge is improved and demonstrated by RAG and knowledge intervention. Extensive experiments and explainability analysis conducted on eight datasets of diverse domains demonstrate the effectiveness of our framework and highlight the collaboration between foundation models and explainable artificial intelligence.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Adrian Bulat and Georgios Tzimiropoulos. 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*, pp. 23232–23241, 2023.

Kathi Canese and Sarah Weis. Pubmed: the bibliographic database. *The NCBI handbook*, 2(1), 2013.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 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*, pp. 178–178. IEEE, 2004.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Janet Hui-wen Hsiao, Hilary Hei Ting Ngai, Luyu Qiu, Yi Yang, and Caleb Chen Cao. Roadmap of designing cognitive metrics for explainable artificial intelligence (xai). *arXiv preprint arXiv:2108.01737*, 2021.

Tim Hulsen. Explainable artificial intelligence (xai): concepts and challenges in healthcare. *AI*, 4 (3):652–666, 2023.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023a.

Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. Guidelines and evaluation of clinical explainable ai in medical image analysis. *Medical image analysis*, 84:102684, 2023b.

Ulf Johansson, Rikard König, and Lars Niklasson. The truth is in there-rule extraction from opaque models using genetic programming. In *FLAIRS*, pp. 658–663, 2004.

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

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 131–138, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36, 2024.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15148–15158, 2023a.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023b.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7086–7096, 2022.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

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

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Mattia Rigotti, Christoph Miksovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *International conference on learning representations*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

Andrea Tocchetti and Marco Brambilla. The role of human knowledge in explainable ai. *Data*, 7 (7):93, 2022.

Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, pp. ocae045, 2024.

Huiling Xiang, Yongjie Xiao, Fang Li, Chunyan Li, Lixian Liu, Tingting Deng, Cuiju Yan, Fengtao Zhou, Xi Wang, Jinjing Ou, et al. Development and validation of an interpretable model integrating multimodal information for improving ovarian cancer diagnosis. *Nature Communications*, 15 (1):2681, 2024.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*, 2024.

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

Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23438–23448, 2024.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. arxiv 2022. *arXiv preprint arXiv:2205.01917*, 2022.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10899–10909, 2023.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024a.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023a.

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13041–13049, 2020.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

APPENDIX FOR "KEEP: TOWARDS A KNOWLEDGE-ENHANCED EXPLAINABLE PROMPTING FRAMEWORK FOR VISION-LANGUAGE MODELS"

A APPENDIX: DATASET DETAILS (WITH GENERATED CONCEPTS FOR MEDICAL DOMAIN)

Derm7pt. *Derm7pt* (Kawahara et al., 2018) is a dermoscopic image dataset containing 1,011 images with clinical concepts for melanoma skin lesions in dermatology. Only the dermoscopic images are considered in this paper. We use the category classification of *normal* and *melanoma*, where the melanoma scores and a threshold $thres = 1$ are used to categorize the images (Kawahara et al., 2018). Clinical concepts for diagnosing melanoma include "Pigment Network", "Dots and Globules", "Pigmentation", "Streaks", "Regression Structures", "Blue-Whitish Veil" and "Vascular Structures".

Pneumonia. The *Pneumonia* dataset (Kermany et al., 2018) is a public dataset for classifying *pneumonia* cases from *normal* ones, which includes 5,863 chest X-ray images. The official dataset splitting is adopted. The clinical concepts for diagnosing pneumonia include "Pleural Effusion", "Infiltration", and "Consolidation".

Open-i. *Open-i* (Demner-Fushman et al., 2016) is a chest X-ray dataset with 3,955 radiology reports, corresponding to 7,470 frontal and lateral images. We filter out the lateral x-ray, leaving only frontal images. Following previous work, we further filter out diseases and leave the three main categories, including *normal*, *opacity*, and *cardiomegaly*. The generated clinical concepts we adopted are "Atelectasis", "Pleural Effusion", "Infiltration", "Consolidation", "Pneumonia", and "Edema".

CCBTM. *CCBTM* (Crystal Clean: Brain Tumors MRI Dataset (Hashemi, 2023)) is a brain tumor MRI dataset containing 21,672 images. The categories cover the main tumor types, including *glioma tumor*, *meningioma tumor*, *pituitary tumor*, and a *normal* class. The dataset is split into training set, validation set, and test set according to the proportion of 70%, 15% and 15%, respectively. The generated clinical concepts for diagnosing brain tumors include "Edema", "Calcifications", and "Infiltration".

Caltech101. The *Caltech101* dataset (Fei-Fei et al., 2004) includes images of generic objects belonging to 101 categories, with about 40 to 800 images per category. We adopt the split following CoOp (Zhou et al., 2022b), where 100 categories are selected with 8,242 images in total, and the numbers of images in the training set, validation set, and testing set are 4,128, 1,649, and 2,465, respectively.

FGVC-Aircraft. The *FGVC-Aircraft* dataset (Maji et al., 2013) contains 10,200 images of aircraft, with 100 images for each of 102 different aircraft model variants, most of which are airplanes. The (main) aircraft in each image is annotated with a tight bounding box and a hierarchical airplane model label. To be consistent with previous works (Zhou et al., 2022b; Gao et al., 2021), 100 categories of aircraft are adopted, and the numbers of images in the training set, validation set, and testing set are 3,334, 3,333, and 3,333, respectively.

Oxford-Flowers102. *Oxford-Flowers102* (Nilsback & Zisserman, 2008) is a natural image dataset for fine-grained classification of flowers, consisting of 102 flower categories with 8189 images in total. Each class consists of between 40 and 258 images. The numbers of images in the training set, validation set, and testing set are 4,093, 1,633, and 2,463, respectively.

DTD. *DTD* (Describable Textures Dataset (Cimpoi et al., 2014)) is a texture datasets containing 5,640 images collected "in the wild" jointly labeled with 47 describable texture attributes (categories). The numbers of images in the training set, validation set, and testing set are 2,820, 1,128, and 1,692, respectively.

B Appendix: Knowledge-Enhanced Prompt Creation Details

B.1 RAG Examples

Figure 7 shows an example of using retrieval-augmented generation for eliciting domain knowledge. The retriever properly retrieves the relevant documents based on the query about pneumonia, which improves the reliability and interpretability of the generated domain knowledge.

Query: What are useful visual features for diagnosing pneumonia in a chest X-ray?

Retrieved Documents:
… evidence of an infiltrate on chest radiography warrants a diagnosis of pneumonia, …
Viral pneumonia characteristically shows diffuse, streaky infiltrates of bronchopneumonia…
The radiographic appearance of pneumococcal pneumonia is varied; it classically consists of lobar or segmental consolidation (Fig. 171-6) but in some cases is patchy …
Bacterial pneumonia characteristically shows lobar consolidation, or a round pneumonia, with pleural effusion in 10% to 30% of cases…
Pleural effusion is an abnormal collection of fluid in the pleural space. It can be seen in pneumonia due to the inflammatory process involving the pleura.

Figure 7: Examples of the retrieval for pneumonia diagnosis. MedCPT (Jin et al., 2023a) is used as the retriever.

B.2 Details of Utilizing Domain-Specific Foundation Models

For disease diagnosis for medical image datasets, we utilized several domain-specific foundation models to generate fine-grained domain knowledge, as illustrated in Section 3.2. Specifically, for chest X-ray images (i.e., *Pneumonia* and *Open-i* datasets), KAD (Zhang et al., 2023b) is adopted for image-wise concept labeling, which leverages existing medical domain knowledge to guide vision-language pre-training using paired chest X-rays and radiology reports. Specifically, to leverage the knowledge of KAD to annotate concept c_i for a given image I with the vision encoder $E_v(\cdot)$ and text encoder $E_t(\cdot)$, we first need to calculate the similarities for image with positive prompt P^{c_i} and negative prompt N^{c_i}:

$$\begin{aligned} sim_p &= E_{\mathrm{DQN}}(E_v(I), E_t(P^{c_i})), \\ sim_n &= E_{\mathrm{DQN}}(E_v(I), E_t(N^{c_i})), \end{aligned} \tag{6}$$

where sim_p and sim_n denote the similarities of the input image with the positive prompt and negative prompt, respectively. $E_{\mathrm{DQN}}(\cdot)$ is an extra proposed disease query network of KAD. Take the concept $c_i =$"pleural effusion" as an example, the positive prompt P^{c_i} is "pleural effusion", while the used negative prompt N^{c_i} is "no pleural effusion". Finally, the absence of concept c_i for image I is decided on the larger one of sim_p and sim_n, for example, if $sim_p > sim_n$, then concept c_i is present in image i (i.e., $Pre_{c_i} = 1$, as mentioned in Section 3.2). In addition, BiomedCLIP (Zhang et al., 2023a) is adopted for brain MRI concept labeling. The way to annotate clinical concepts is almost the same as using KAD except that BiomedCLIP only uses the vision and text encoders without the disease query network. The positive and negative prompts we used in BiomedCLIP for brain tumor concept labeling are "[concept name] presented in this image" and "this is an image of a normal brain", respectively.

B.3 More Knowledge-Enhanced Prompt Examples

More image samples and their corresponding generated knowledge-enhanced prompts are shown in Figure 8.

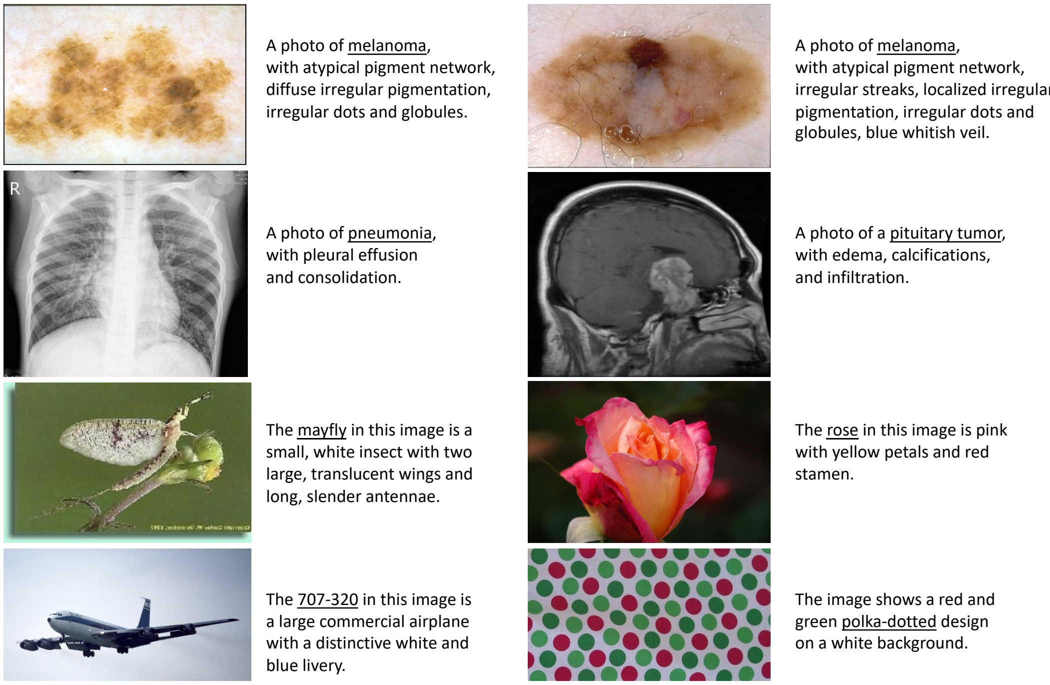
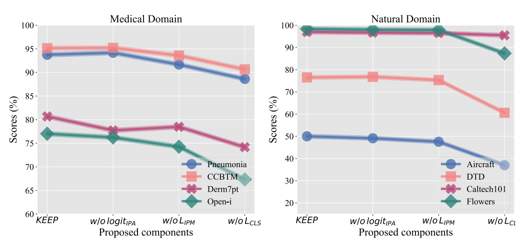


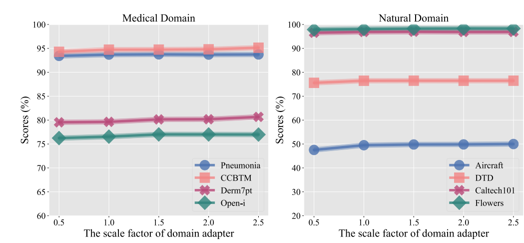
Figure 8: More examples of the images from different domains and their corresponding generated knowledge-enhanced prompts. The category name of each image is underlined.

C Appendix: More Ablation Study Results

More ablation study results are shown in Figure 9. Specifically, we display the complete version of ablation for fusion logits and losses in the medical domain and natural domain in Figure 9(a), which demonstrates the effectiveness of our proposed components. Moreover, ablation results of the scale factors of the domain adapter are presented in Figure 9(b). It can be observed that the overall performance increases and gets stable when the scale factor increases. Since the domain adapter is a learnable matrix that adds to the original text embeddings, a greater scale factor means learning more from the specific domain, where the results are in line with expectations.



(a) Ablation study of the fusion logits and losses.



(b) Ablation of the scale factors of the domain adapter.

Figure 9: Ablation study results. (a) The complete ablation study of fusion logits and losses, the detailed version of Table 4. (b) The ablation study of the scale factors of the domain adapters for each dataset from various domains.

D Appendix: Computational Efficiency

To evaluate the computational efficiency of our method, we report the training and inference compute cost in Table 5. The results demonstrate that our method achieves the best model performance while showing promising computational efficiency, with the best inference time and FPS compared to CoCoOp (Zhou et al., 2022a) and LASP (Bulat & Tzimiropoulos, 2023).

Table 5: Computational efficiency comparison using *Penumonia* dataset. Evaluation of average training (per epoch) and inference time (second) for all methods is conducted on a single RTX4090 GPU. PERFORMANCE is the average classification accuracy on eight considered datasets.

METHOD	TRAINING TIME ↓	INFERENCE TIME↓	FPS ↑	PERFORMANCE↑
CoCoOp	109.21	5.62	121	76.14
LASP	20.38	0.86	732	79.32
KEEP	22.82	0.72	912	83.55