XCoOp: Explainable Prompt Learning for Computer-Aided Diagnosis via Concept-guided Context Optimization

Yequan Bie[1], Luyang Luo[1], Zhixuan Chen[1], and Hao Chen[1,2]✉

[1]The Hong Kong University of Science and Technology, Hong Kong, China
[2]HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen, China
{ybie, zchenhi}@connect.ust.hk, cseluyang@ust.hk, jhc@cse.ust.hk

Abstract. Utilizing potent representations of the large vision-language models (VLMs) to accomplish various downstream tasks has attracted increasing attention. Within this research field, soft prompt learning has become a representative approach for efficiently adapting VLMs such as CLIP, to tasks like image classification. However, most existing prompt learning methods learn text tokens that are unexplainable, which cannot satisfy the stringent interpretability requirements of Explainable Artificial Intelligence (XAI) in high-stakes scenarios like healthcare. To address this issue, we propose a novel explainable prompt learning framework that leverages medical knowledge by aligning the semantics of images, learnable prompts, and clinical concept-driven prompts at multiple granularities. Moreover, our framework addresses the lack of valuable concept annotations by eliciting knowledge from large language models and offers both visual and textual explanations for the prompts. Extensive experiments and explainability analyses conducted on various datasets, with and without concept labels, demonstrate that our method simultaneously achieves superior diagnostic performance, flexibility, and interpretability, shedding light on the effectiveness of foundation models in facilitating XAI. The code is available at https://github.com/Tommy-Bie/XCoOp.

Keywords: Prompt Learning · XAI · Multi-modal ML · LLM · VLM

1 Introduction

In the era of foundation models (FMs), large-scale vision-language pre-trained models (VLMs) such as CLIP [24], BLIP [18], Flamingo [2], ALIGN [13], CoCa [29] have underscored their potential in representation learning, excelling at vision and language understanding. However, the massive sizes and expensive training costs have prompted studies to explore ways to efficiently adapt the knowledge of pre-trained VLMs to downstream tasks. Recently, prompt learning

✉ Corresponding author.

from the field of natural language processing has been introduced to the vision domain [31, 30], achieving great success in adapting large-scale VLMs to downstream tasks like image classification and segmentation [21, 19]. These methods fix the parameters of the models and train the learnable tokens that serve as the input of the text encoder (i.e., context optimization), significantly reducing the cost of utilizing foundation models. Nevertheless, existing prompt learning methods result in unexplainable learned tokens. This lack of interpretability prevents further application of prompt learning from being applied to high-stakes domains with rigorous demands of trustworthiness, such as healthcare [20, 27, 22, 6]. Specifically, the models applied to the healthcare domain should not only perform well but also need to be understandable and trustworthy to practitioners, necessitating research into Explainable Artificial Intelligence (XAI). Several prior studies have introduced knowledge to prompt learning [28, 4]. For instance, Yao et al. [28] adopt human knowledge (*a photo of a* [*class name*]) as hard prompts to guide the learning of soft prompts at the global level. However, the insufficient knowledge and inadequate guidance still lead to non-interpretable prompt learning. To address the explainability challenge of current methods, we propose **XCoOp**, a novel e**X**plainable prompt learning framework for medical image analysis via concept-guided **Co**ntext **Op**timization, which leverages medical knowledge by aligning the semantics of the images, learnable prompts, and clinical concept-driven prompts at multiple granularities, making each token of soft prompts more informative and explainable guided by clinical concepts of corresponding diseases. Furthermore, our framework addresses the lack of valuable concept annotations by eliciting knowledge from large language models and offers both visual and textual explanations for learned prompts.

We summarize our main contributions as follows: (i) We propose XCoOp, a novel explainable prompt learning framework that leverages concept-based medical knowledge at multiple granularities to make the prompts more explainable. To the best of our knowledge, this is the first work to explore addressing the lack of interpretability of prompt learning methods in healthcare. (ii) We demonstrate that our method can be flexibly applied to various datasets with or without concept annotations, alleviating the requirement of human labor by eliciting medical knowledge from LLMs. (iii) Extensive experiments and explainability analyses show that our method simultaneously achieves promising performance and interpretability, highlighting the effectiveness of foundation model-enhanced XAI.

2 Method

Fig. 1 presents the overall architecture of our explainable prompt learning framework for computer-aided diagnosis. Specifically, we initialize the soft prompts with "*a photo of a* [*disease name*]", and the clinical prompts are created based on the medical concepts (Section 2.1). The text features of the prompts are extracted using a pre-trained text encoder, and a multi-granularity prompt alignment module is proposed to align the semantics of the soft prompts and the clinical prompts (Section 2.2). The final disease diagnosis is performed by mea-

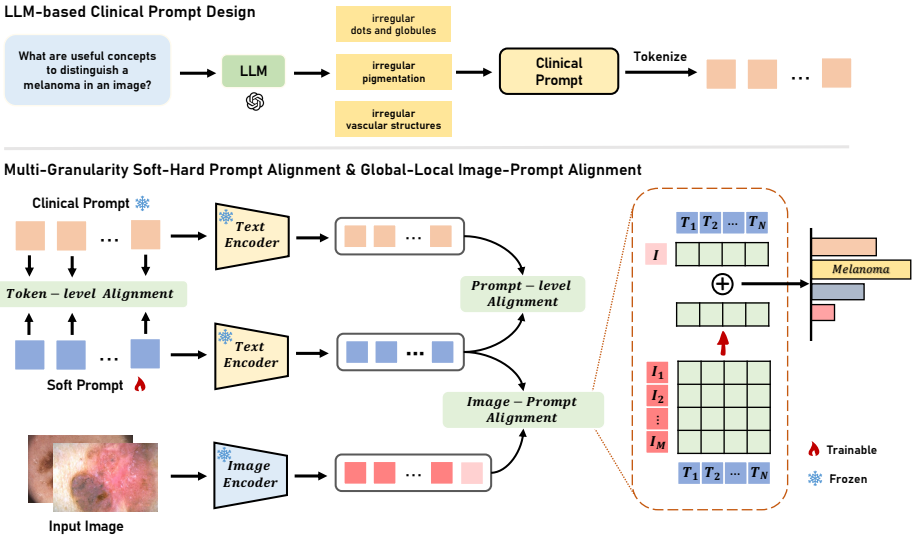


Fig. 1. The overall pipeline of XCoOp. The key insight of XCoOp is enhancing the informativeness and explainability of the soft prompts under the guidance of concept-based medical knowledge at multiple granularities, achieving FM-enhanced XAI.

suring the similarity between the text features of soft prompts and the image features at both global and local levels (Section 2.3).

2.1 Clinical Concept-Driven Prompt Design

To introduce medical knowledge into the prompt learning process, we first design disease-specific prompts using clinical concepts. Fig. 1 shows the steps of creating clinical prompts in our framework. Specifically, for medical datasets with concept annotations (e.g., Derm7pt [15], SkinCon [7]), we can easily create clinical prompts based on the labels annotated by medical experts. An example clinical prompt for melanoma in a dermoscopic image is *a photo of melanoma, with irregular pigment network, dots and globules, blue-whitish veil, and vascular structures*. For the datasets lacking explicit concept annotations, we elicit medical knowledge from a large language model such as GPT4 [1] and create the corresponding clinical prompts. A sample query used to prompt the LLM is *"What are the most useful visual concepts to distinguish* [*disease name*] *in a* {*dermoscopic image, chest X-ray, etc.*}*?"*.

2.2 Soft-Hard Prompt Alignment

To enhance the informativeness and explainability of the soft prompts by incorporating clinical semantics, we introduce a soft-hard prompt alignment module

that aligns prompts at both the prompt level and token level. Prompt-level alignment facilitates the model to learn correspondences between soft prompts and clinical (hard) prompts from a global disease perspective, exploiting the intrinsic information captured by the pre-trained text encoder. The token-level alignment focuses on a more fine-grained local level. Since each token embedding of the clinical prompts is obtained by tokenizing the original concept-based prompts, the alignment enforces the soft prompts to be close to the clinical prompts in the embedding space, aiming to make each token of soft prompts more informative and explainable guiding by clinical concepts of corresponding diseases, hence enhancing the effectiveness and interpretability of the prompt learning framework.

Token-level Alignment. Given the token embeddings of soft prompts $V \in \mathbb{R}^{D\times C\times dim}$ and clinical prompts $Q \in \mathbb{R}^{D\times C\times dim}$ for different classes, we first align their embeddings at the token level via contrastive learning, where D, C, dim represent the number of classes, the context length, and dimension of embedding, respectively. A probability distribution over the class labels is given by:

$$P(y_d|V_d) = \frac{\exp(\cos(Q_d, V_d)/\tau)}{\sum_{k=1}^{D}\exp(\cos(Q_k, V_d)/\tau)}, \tag{1}$$

where y_d is the binary label of class d, $\cos(\cdot,\cdot)$ is the cosine similarity, and τ is a temperature parameter. The token-level alignment objective $\mathcal{L}_T$ is optimized by minimizing the cross-entropy loss:

$$\mathcal{L}_T = -\sum_{k=1}^{D} y_k \log P(y_k|V_d). \tag{2}$$

Prompt-level Alignment. Given the pre-trained text encoder $g(\cdot)$, we align the text features of soft prompts and clinical prompts at the global prompt level by minimizing the following objective function:

$$\mathcal{L}_P = \sum_{d=1}^{D} CE\left(\frac{\exp(\cos(g(Q_d), g(V_d)))/\tau)}{\sum_{k=1}^{D}\exp(\cos(g(Q_k), g(V_d))/\tau)}, y_d\right), \tag{3}$$

where $\mathcal{L}_P$ represents the prompt-level alignment loss, $CE(\cdot)$ denotes the cross-entropy loss, $g(V)$ and $g(Q) \in \mathbb{R}^{D\times dim}$ denote the output text features of soft prompts V and clinical prompts Q, respectively. The overall objective of the soft-hard prompt alignment module $\mathcal{L}_{\text{PPA}}$ is the average of $\mathcal{L}_T$ and $\mathcal{L}_P$.

2.3 Global-Local Image-Prompt Alignment

Medical diagnosis typically hinges on various clinical symptoms observable within specific, localized regions in an image. Given that different clinical concepts may correspond to distinct sub-regions of a medical image, we employ a global-local image-prompt alignment module to align the medical images and clinical concept-driven prompts at multiple levels. Specifically, as illustrated in Fig. 1,

given an image x and the pre-trained image encoder of CLIP [24], we obtain the global visual feature p and a set of local features $F = \{f_1, f_2, ..., f_M\}$, where M is the number of local (patch) features. The final prediction probability is computed by the matching scores of both global and local features, and the alignment can be optimized using cross-entropy loss which estimates the discrepancy between the predicted diagnosis results and the ground truth:

$$\mathcal{L}_{\text{IPA}} = CE[\cos(p, g(V_d)) + \lambda \frac{1}{M}(\sum_{m=1}^{M} \cos(f_m, g(V_d))), y_d], \tag{4}$$

where $\mathcal{L}_{\text{IPA}}$ represents the image-prompt alignment loss, λ is the weight of the prediction of local features. The overall training objective is represented as $\mathcal{L} = \mathcal{L}_{\text{PPA}} + \lambda' \mathcal{L}_{\text{IPA}}$, where λ' is a loss-balancing factor. The global-local image-prompt alignment module encourages the model to mimic the process wherein medical experts utilize both global and local information to diagnose disease.

3 Experiments

3.1 Experimental Setup

Datasets: ***Derm7pt*** [15] is a dermoscopic image dataset contains 1011 images with clinical concepts for melanoma skin lesions in dermatology. ***SkinCon*** [7] is a skin disease dataset densely annotated by experts for fine-grained model debugging and analysis. The concepts of *Derm7pt* and the *F17k* part of *SkinCon* are used to design clinical prompts for these two datasets. ***Pneumonia*** [16] is a public dataset for classifying pneumonia cases from normal ones, with 1583 normal and 4273 pneumonia images. ***IU X-Ray*** [8] 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.
Implementation Details: Our framework adopted the pre-trained visual (ViT-B/16) and text encoder of CLIP [24]. We adopted SGD [26] optimizer with learning rate of 0.032. We used warm-up and cosine anneal as training strategies. All methods implemented in this paper adopted random crop and random flip for data augmentation. Grid search was used to select hyperparameters, we set $\tau = 0.9, \lambda = 0.1$. All experiments were conducted on an RTX 3090 GPU.

3.2 Experimental Results

In order to comprehensively demonstrate the competitive performance of our method in disease diagnosis, we commence by comparing with other state-of-the-art prompt learning methods on four datasets. Subsequently, we undertake intensive ablation experiments to assess the effectiveness of our method. Finally, we evaluate the explainability of our framework using multiple XAI metrics.

Table 1. Quantitative comparison on disease diagnosis with the state-of-the-art prompt learning methods. The performance is reported as mean_{std} of three random runs. Our method is highlighted in light cyan, and the best results are shown in **bold**.

METHOD	Derm7pt		SkinCon		Pneumonia		IU X-Ray	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
CLIP [24]	50.00	69.11	39.68	70.29	50.00	62.52	47.90	13.21
CoOp [31]	$71.76_{0.1}$	$75.19_{0.4}$	$77.52_{0.4}$	$75.91_{0.7}$	$84.08_{0.6}$	$85.88_{0.6}$	$78.45_{1.2}$	$71.93_{0.7}$
CoCoOp [30]	$70.40_{0.4}$	$77.04_{0.7}$	$78.02_{0.5}$	$76.19_{0.8}$	$85.96_{0.4}$	$86.06_{0.8}$	$76.00_{1.6}$	$70.63_{0.5}$
KgCoOp [28]	$69.67_{2.7}$	$73.84_{1.4}$	$75.33_{0.3}$	$76.95_{0.5}$	$80.95_{0.3}$	$82.64_{0.3}$	$75.61_{1.2}$	$70.74_{1.2}$
LASP [4]	$75.08_{0.6}$	$76.20_{1.6}$	$78.31_{0.3}$	$77.33_{0.8}$	$91.31_{0.1}$	$92.41_{0.1}$	$83.69_{0.3}$	$76.46_{0.7}$
XCoOp	$\mathbf{78.43}_{0.6}$	$\mathbf{78.82}_{1.0}$	$\mathbf{81.12}_{0.3}$	$\mathbf{78.57}_{0.6}$	$\mathbf{92.85}_{0.3}$	$\mathbf{93.80}_{0.3}$	$\mathbf{84.91}_{0.6}$	$\mathbf{78.44}_{0.9}$

Diagnosis Results. In Table 1, we report the disease diagnosis comparison results of our method using AUROC and Accuracy on four medical image datasets. We include the CLIP baseline [24] without any tuning (the first row), two CoOp-based methods (CoOp [31] and CoCoOp [30]), and two knowledge-guided prompt learning methods (KgCoOp [28] and LASP [4]). Our method outperforms other state-of-the-art prompt learning methods with a significant margin, especially achieving 1.2% $\sim$ 3.4% AUC and 1.2% $\sim$ 2.0% accuracy improvement compared to the second-best results on all considered datasets, which demonstrates that the full utilization of medical knowledge and the global-local correlations between images and prompts effectively encourages the soft prompts to learn clinical semantics, thus benefiting the performance of our model.

Ablation Study. We conduct various ablation studies on *SkinCon* to investigate the influence of different modules and settings. In Table 2, we assess the effectiveness of each module in our proposed framework. Specifically, we show that our method can benefit from all the components, including the clinical

Table 2. Ablation study of XCoOp on disease diagnosis (AUC [%]). CCP, IPA, and PPA represent the clinical concept-driven prompts, image-prompt alignment, and soft-hard prompt alignment modules, respectively.

Method	AUC
Baseline (LASP [4])	$78.31_{0.3}$
CCP	$79.93_{0.4}$
CCP + IPA	$80.46_{0.7}$
CCP + IPA + PPA	$\mathbf{81.12}_{0.3}$

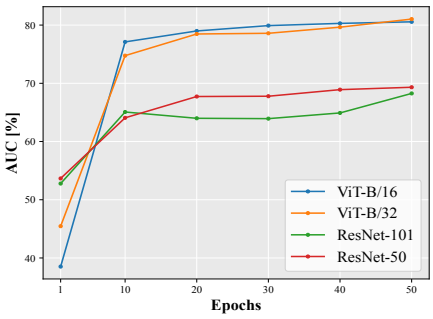
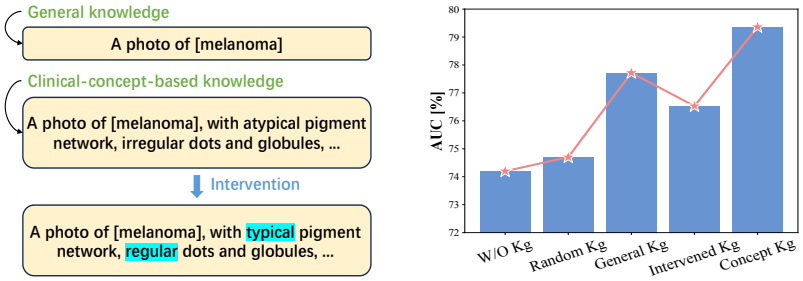


Fig. 2. Ablation study on the number of training epochs of XCoOp with different vision backbones.



(a) Prompt examples. (b) Knowledge intervention.

Fig. 3. Illustration of our model's faithfulness using knowledge intervention. (a) Clarification of prompt examples based on different knowledge and intervention. (b) The results of concept-based knowledge (Kg) intervention on Derm7pt, where x-axis represents different kinds of prompts and y-axis represents the AUC [%], respectively.

concept-driven prompts, the soft-hard prompt alignment, and the global-local image-prompt alignment. The last configuration of Table 2 demonstrates that our method achieves the best overall performance with all designed components. To explore the influence of different numbers of training epochs and vision backbones, we conduct an ablation study and report the AUC in Fig. 2. The results show that our method can converge quickly with different vision backbones (e.g., ViT [9], ResNet [11]), which demonstrates the high efficiency of our method.

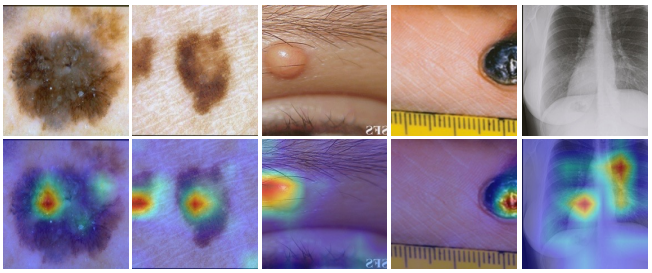
3.3 Analysis of Explainability

In order to evaluate the explainability of our proposed method, we analyze our framework using multiple crucial XAI metrics in this section. Specifically, inspired by previous works [10, 25, 12, 3, 14], we evaluate our framework from the perspectives of *faithfulness*, *understandability* and *plausibility*.

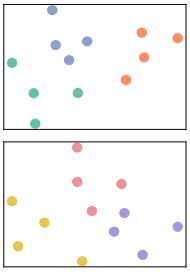
Faithfulness. *Faithfulness* is defined as the degree to which the explanation reflects the model decision process and requires the explanation to be faithful to the designed model mechanism [10, 17, 25]. In this paper, we evaluate *faithfulness* by intervening the input clinical concept-driven prompts. As shown in Fig. 3, we use five kinds of prompt settings, including without knowledge, with random knowledge (i.e., random words as clinical prompts), with general knowledge (i.e., knowledge without specific clinical concepts), with clinical-concept-based knowledge and the intervened knowledge. Specifically, we adopt *Derm7pt* dataset as an example, as shown in Fig. 3(a), where intervention means modifying some of the concepts in the original clinical prompts and obtaining a new prompt. Fig. 3(b) shows that using only random knowledge, general knowledge, or knowledge after intervention as prompts may lead to performance degradation, which demonstrates that the clinical knowledge faithfully explains the model's decisions.

Table 3. Quantitative comparison on prompt interpretation by measuring distances between the soft prompts and the hand-crafted clinical prompts (i.e., textual explanations). The results are reported as the average distances of different categories. Our method is highlighted in light cyan, and the best results are shown in **bold**.

Method	Derm7pt	SkinCon	Pneumonia	IU X-ray	**Average** ↓
KgCoOp [28]	2.293	1.475	1.727	2.433	1.982
LASP [4]	2.936	3.867	2.270	2.972	3.011
XCoOp	**1.161**	**1.139**	**0.987**	**1.127**	**1.104**



(a) Image-prompt similarity visualization.



(b) t-SNE.

Fig. 4. Visual explanations. (a) Visualization examples of the similarity between the images and soft prompts. (b) The t-SNE visualization of tokens of different soft prompts of SkinCon (top) and IU X-ray (bottom) datasets. Different colors represent different categories of prompts, and the number of context tokens is 4.

Understandability & Plausibility. *Understandability* requires explanations to be easily understandable by users without requiring technical knowledge [14] while *plausibility* refers to given domain knowledge, how believable or likely the explanation seems [10, 5]. Our framework achieves *understandability* and *plausibility* by offering both textual and visual explanations. Specifically, we interpret the learnable prompts by measuring the distance between the soft prompts and the hand-crafted clinical prompts. As shown in Table 3, we compare the average distances with two knowledge-guided prompt learning methods [28, 4]. Our method outperforms the other methods and achieves the least distance between the learnable prompts and the clinical prompts. For visual explanation, our framework provides the similarity visualization between the medical images and the learnable prompts, as shown in Fig. 4(a), where we can observe that the model focuses more on discriminative concept regions within images guided by our learned prompts. Fig. 4(b) presents the t-SNE visualization [23] of tokens of different soft prompts and shows that the token embeddings cluster well, demonstrating that tokens in each prompt meticulously learn the discriminative clinical semantics of the corresponding disease category. The explanations offered by our framework enhance human comprehension of the model's decision-making process by elucidating the utilized knowledge and the specific regions of focus, potentially aiding medical experts in utilizing AI models for disease diagnosis.

4 Conclusion

In this paper, we propose XCoOp, an explainable prompt learning framework for computer-aided diagnosis, which utilizes medical knowledge by aligning the semantics of images, learnable prompts, and clinical concept-driven prompts at multiple granularities. By adopting the concept-based knowledge eliciting from foundation models to guide the soft prompt at both the token embedding level and prompt level, our method outperforms other prompt learning methods while preserving inherent interpretability with both visual and textual explanations. Extensive experiments and explainability analyses conducted on various datasets demonstrate that our method simultaneously achieves promising performance and interpretability, highlighting the effectiveness of FM-enhanced XAI.

Acknowledgments. This work was supported by the HKUST (Project No. FS111) and Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083).

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
3. Bie, Y., Luo, L., Chen, H.: Mica: Towards explainable skin lesion diagnosis via multi-level image-concept alignment. arXiv preprint arXiv:2401.08527 (2024)
4. Bulat, A., Tzimiropoulos, G.: 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)
5. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)
6. Chen, Z., Luo, L., Bie, Y., Chen, H.: Dia-llama: Towards large language model-driven ct report generation. arXiv preprint arXiv:2403.16386 (2024)
7. Daneshjou, R., Yuksekgonul, M., Cai, Z.R., Novoa, R., Zou, J.Y.: Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. Advances in Neural Information Processing Systems **35**, 18157–18167 (2022)
8. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)

9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hsiao, J.H.w., Ngai, H.H.T., Qiu, L., Yang, Y., Cao, C.C.: Roadmap of designing cognitive metrics for explainable artificial intelligence (xai). arXiv preprint arXiv:2108.01737 (2021)
13. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
14. Jin, W., Li, X., Fatehi, M., Hamarneh, G.: Guidelines and evaluation of clinical explainable ai in medical image analysis. Medical Image Analysis **84**, 102684 (2023)
15. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics **23**(2), 538–546 (2018)
16. 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.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
17. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: 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)
18. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
19. Lin, Y., Nie, D., Liu, Y., Yang, M., Zhang, D., Wen, X.: Multi-target domain adaptation with prompt learning for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 717–727. Springer (2023)
20. Lipton, Z.C.: The doctor just won't accept that! arXiv preprint arXiv:1711.08037 (2017)
21. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7086–7096 (2022)
22. Luo, L., Huang, X., Wang, M., Wan, Z., Chen, H.: Medical image debiasing by learning adaptive agreement from a biased council. arXiv preprint arXiv:2401.11713 (2024)
23. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

25. Rigotti, M., Miksovic, C., Giurgiu, I., Gschwind, T., Scotton, P.: Attention-based interpretability with concept transformers. In: International Conference on Learning Representations (2021)
26. Robbins, H., Monro, S.: A stochastic approximation method. The annals of mathematical statistics pp. 400–407 (1951)
27. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
28. Yao, H., Zhang, R., Xu, C.: 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)
29. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arxiv 2022. arXiv preprint arXiv:2205.01917
30. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
31. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)