

Efficient In-Context Medical Segmentation with Meta-driven Visual Prompt Selection

Chenwei Wu¹, David Restrepo², Zitao Shuai¹,
Zhongming Liu¹, and Liyue Shen¹ ✉

¹ University of Michigan, Ann Arbor, United States

² Massachusetts Institute of Technology, Cambridge, United States
✉liyues@umich.edu

Abstract. In-context learning (ICL) with Large Vision Models (LVMs) presents a promising avenue in medical image segmentation by reducing the reliance on extensive labeling. However, the ICL performance of LVMs highly depends on the choices of visual prompts and suffers from domain shifts. While existing works leveraging LVMs for medical tasks have focused mainly on *model-centric* approaches like finetuning, we study an orthogonal *data-centric* perspective on how to select good visual prompts to facilitate generalization to medical domain. In this work, we propose a label-efficient in-context medical segmentation method by introducing a novel *Meta-driven Visual Prompt Selection* mechanism (MVPS), where a prompt retriever obtained from a meta-learning framework actively selects the optimal images as prompts to promote model performance and generalizability. Evaluated on 8 datasets and 4 tasks across 3 medical imaging modalities, our proposed approach demonstrates consistent gains over existing methods under different scenarios, improving both computational and label efficiency. Finally, we show that MVPS is a flexible, finetuning-free module that could be easily plugged into different backbones and combined with other model-centric approaches.

Keywords: In-Context Learning · Meta-Learning · Active Learning · Reinforcement Learning · Medical Image Segmentation.

1 Introduction

Large vision models (LVMs), such as SegGPT [27], Painter [26], and variants of SAM [12], can learn to perform segmentation on unseen tasks from a few visual prompts without updating model parameters. This in-context learning (ICL) ability enables label-efficient medical image segmentation [29, 5] without relying on expensive expert annotations and thus ease the deployment of medical AI.

However, the effectiveness of ICL depends on the quality and number of visual prompts [24, 32]. Transferring LVMs from natural images to medical images faces further challenges due to domain shifts [5] and the large variability across patient populations, demographics, imaging protocols, etc. For example,

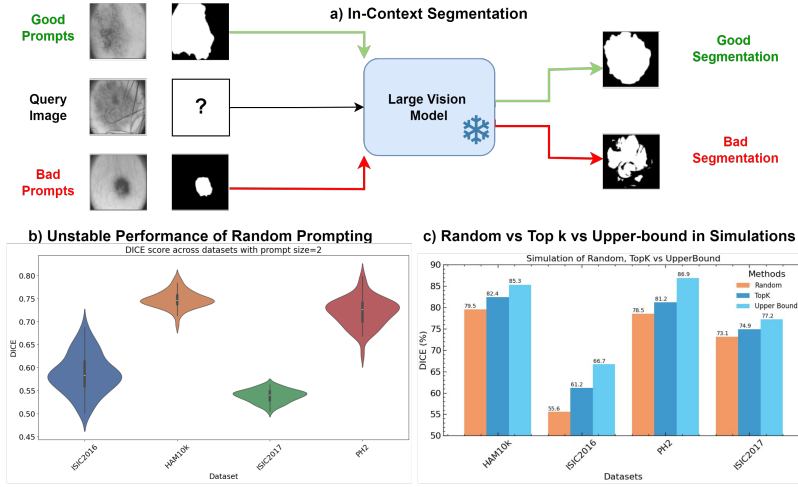


Fig. 1: **a) In-context segmentation.** Large Vision Models are capable of taking in visual prompts of image-mask pairs and outputting the segmentation mask prediction for the query image. **b) Instability of ICL with random prompting.** ICL has shown unstable performance using random prompts. By conducting experiments of in-context skin lesion segmentation using SegGPT [27] on 4 different dermatology datasets [25, 4, 17], with prompt size of 2, the results show a large variance. (Mean DICE scores: 37.36%, 14.84%, 11.46%, and 31.80%). **c) Better prompts lead to a significant improvement in ICL.** In this simulation study, we iterate through all prompt selection options (with prompt size of 2 in this example) given a prompt pool of 100 images and test ICL performance. There is plenty of room for improvement over the current prompt selection methods like TopK [32] approach.

Figure 1(b) demonstrates this issue in dermatology image segmentation. Using an LVM with two randomly selected visual prompts, the performance varies by 20.11% on average across four different dermatology image datasets [25, 4, 17, 8]. ICL also demonstrates high sensitivity to the number of prompts in challenging segmentation tasks such as vessel segmentation.

While the critical role of prompt selection has gained significant attention in the Large Language Model (LLM) field [31], LVM adaptation works have been focusing primarily on the model-centric approaches such as efficient model fine-tuning [10, 33]. Some works [21, 14] have explored the impacts of prompt quality on LLM adaptation to new data and tasks; [16, 31] focuses on formulating the selective prompt labeling as a data-centric method to determine which prompt samples are valuable for labeling, similar to active learning. On the imaging side, [15] adapts LVMs to medical images by fine-tuning the model on extensive medical images, but the complexity of training large-scale models and the need for massive and expensive labeled data highlight the significant resources

required for such comprehensive models; [29] utilizes an adaptor to fine-tune LVMs efficiently to medical datasets but remains model-centric. [32] is the first work to discuss visual prompt selection on natural images and proposes an unsupervised TopK approach and a supervised method SupPR for prompt selection. SupPR requires training of a large visual embedding extractor and demonstrating slight improvements over TopK approach. Other works have also explored how to improve the prompt quality given a fixed set of prompts: [24] proposes prompt fusion, using an ensemble of different prompt layouts to activate knowledge at different positions in the LVMs; [30] devises a prompt enhancer trained on large-scale labeled data to add perturbations to the in-context prompt pairs. However, these works still rely on TopK prompt selection, which are constrained by the distance measurements and cannot always find the visual prompt optimal across domains. Overall, there hasn't been a generic solution to optimizing visual prompt selection to adapt LVMs efficiently for in-context medical segmentation.

Herein, we introduce MVPS, a novel Meta-learning-driven Visual Prompt Selection framework. The key contributions of our research lie in three-fold:

1. We propose a label-efficient in-context medical segmentation method enabled by introducing a novel Meta-driven Visual Prompt Selection mechanism (MVPS), which enables cross-domain adaptation of LVMs to medical imaging.
2. Specifically, we propose a meta-driven active visual prompt retrieval approach, by constructing a meta-learning scheme to teach a transformer-based prompt retriever which images are worthy of being selected as visual prompts to boost model performance. The prompt retriever is optimized through the probability distribution estimation and reward policy gradients.
3. MVPS provides a data-centric and finetuning-free enhancement to in-context medical segmentation, with both data and label efficiencies largely benefiting medical applications. MVPS consistently outperforms baselines across 8 datasets and 4 tasks across 3 medical imaging modalities. We also show that MVPS is flexible enough to plug into different LVM backbones, and to be combined with model-centric approaches.

2 Methodology

In the following section, we will introduce the technical details of the proposed MVPS framework as shown in Fig 2. Taking Dermatology as an example, given a training set (e.g., HAM10k) and a testing set (e.g., ISIC), the MVPS framework constructs tasks for meta-training and meta-testing. Each task consists of a **Support Set** of 1000 unlabeled images and a **Query Set** of 100 labeled image-mask pairs. The goal of meta-training is to learn a prompt retriever $f(\cdot)$ that can select the best prompts from the support set to enhance segmentation performance on the query set. The retriever is optimized by comparing the output segmentation masks of the Large Vision Model (LVM) against the ground truth masks in the query set, and then using this feedback to improve prompt selection. For meta-testing, the retriever selects prompts from a simulated prompt

pool of 1000 images in the test dataset. The LVM uses these prompts to perform segmentation on new query images. During meta-testing, we evaluate and report the LVM’s segmentation performance on different target (meta-test) datasets in order to mimic the real-world scenario when the LVMs are utilized in a different hospital with a new patient cohort.

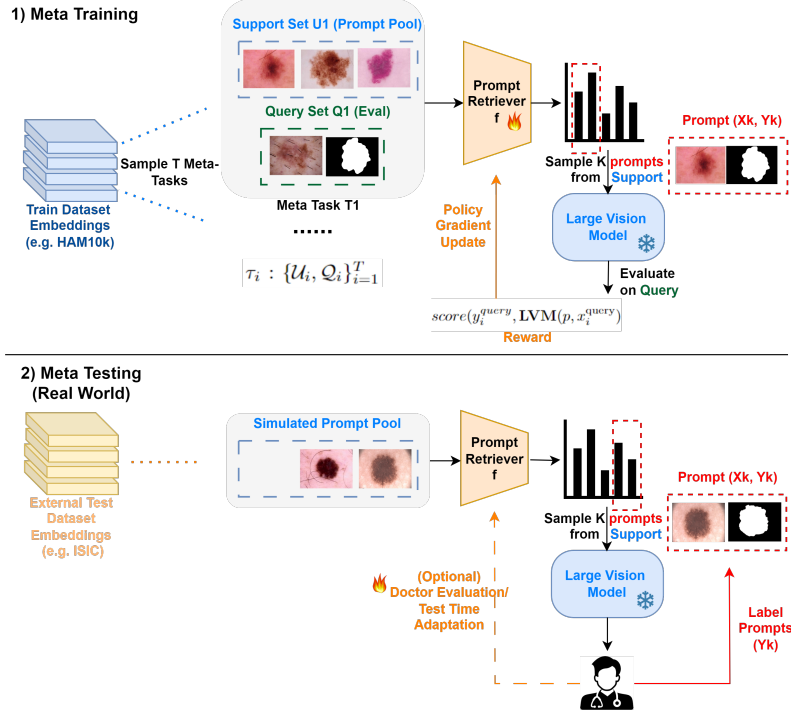


Fig. 2: Meta-training and meta-testing stages of the proposed MVPS framework (use dermatology dataset as an example). Note that prompt retriever is trainable while large vision model is kept frozen.

Meta Tasks Construction. As shown in Fig. 2, in our meta-learning setting, each task is an in-context medical segmentation problem, denoted as $\tau_i : \{\mathcal{U}_i, \mathcal{Q}_i\}_{i=1}^T$. Suppose there are a total of T tasks. Each task consists of an unlabeled image set $\mathcal{U} = \{x_j\}_{j=1}^{1000}$, which we refer to as the support set or prompt pool, and a labeled image set $\mathcal{Q} = \{(x_k, y_k)\}_{k=1}^{100}$ which we refer to as query set consisting of a set of image x and segmentation label y . Both $\mathcal{Q}$ and $\mathcal{U}$ are sampled from the training datasets and there is no overlap between the images in $\mathcal{Q}$ and $\mathcal{U}$ within a single task. In this way, we define each meta prompt learning task τ_i as: $\tau_i = (\mathcal{U}_{\tau_i}^{\text{support}}, \mathcal{Q}_{\tau_i}^{\text{query}})$, where $\mathcal{U}_{\tau_i}^{\text{support}}$ indicates the support set as

the candidate pool for prompt selection, and $\mathcal{Q}_{\tau_i}^{\text{query}}$ indicates the query set to evaluate the in-context learning performance for the current task τ_i . Following the setup of active learning, we hold out specific disease categories if disease category labels are available. We only sample the task τ_i from the non-heldout labels for the meta-training tasks. For meta-validation, we sample from a mix of non-heldout labels and held-out labels. This ensures that meta-training tasks and meta-validation tasks are not sampled from too similar domains to prevent the model from simply memorizing training samples.

Prompt Retriever. Following previous active learning research [20], we design a trainable prompt retriever using the transformer architecture. The retriever takes in a pair of extracted embeddings for both unlabeled and labeled images ($\mathcal{U}_{\tau_i}^{\text{support}}, \mathcal{Q}_{\tau_i}^{\text{query}}$) and outputs a softmax-normalized distribution which indicates the probability of all the samples in the prompt pool to be selected as a good prompt over the prompt pool samples. In order to distinguish the support and query representations [7], we add two learnable vectors that represent prompt pool and query sets, respectively, to their corresponding embeddings. Based upon this predicted distribution, we select k samples with the highest probabilities, along with their labels, as the visual prompts. These image-mask pairs will serve as inputs to the LVMs for completing in-context segmentation tasks to predict the segmentation masks for query images.

Task Augmentation for Better Generalization. To perform more robust meta-training and address the data scarcity problem in domains with small training sets, we propose a new task augmentation by mixing up images and segmentation masks. Specifically, we extend the idea of mix-ups to task augmentations to further densify the task distribution. Randomly selecting an anchor task τ_i , and another task τ_j , we form a new task and support set by interpolating the $(x_{\tau_i}^{\text{support}}, x_{\tau_j}^{\text{support}})$ and corresponding masks $(y_{\tau_i}^{\text{support}}, y_{\tau_j}^{\text{support}})$ from τ_i and τ_j , using a mixing ratio $\lambda \in [0, 1]$, so that $\tilde{x}_{\tau_i} = (1 - \lambda)x_{\tau_i} + \lambda x_{\tau_j}$, $\tilde{y}_{\tau_i} = (1 - \lambda)y_{\tau_i} + \lambda y_{\tau_j}$.

Optimization and Reward Shaping. The reward $\mathcal{R}$ is determined by a scoring function that evaluates the efficacy of the Large Vision Model (LVM) on the query set $\mathcal{Q}_{\tau_i}$, when given a selected prompt p from $\mathcal{U}$, formally denoted as

$$\mathcal{R}(p, \mathcal{Q}_{\tau_i}^{\text{query}}) = \text{score}(y_i^{\text{query}}, \mathbf{LVM}(p, x_i^{\text{query}})), x_i, y_i \in \mathcal{Q}_{\tau_i}^{\text{query}} \quad (1)$$

Here, we employ the DICE coefficient on the query dataset to calculate this reward. The aim is to train the prompt retriever to act as a policy model to discern and generalize the characteristics of good contextual prompts and enable the selection of informative unlabeled examples in scenarios where the retriever may not have exposure to any labeled instances. We utilize the simple but effective expected reinforcement performance as our objective [28, 31, 22] and optimize by

back-propagation of policy-gradients

$$\nabla_{\theta} \mathcal{L} = \mathbb{E}_{\sim \mathbb{P}_{\theta}(p | \mathcal{U}_{\tau_i}^{\text{support}})} [(\mathcal{R}(p, \mathcal{Q}_{\tau_i}^{\text{query}})) \nabla_{\theta} \log(\mathbb{P}_{\theta}(p | \mathcal{U}_{\tau_i}^{\text{support}}))] \quad (2)$$

Here we approximate the expected value by a single Monte-Carlo sample. Previous works in reinforcement learning [31, 2] have found that sparse reward learning is unstable and difficult. Reshaping the reward function into an intuitive additional gain of getting the retrieved sample will help stabilize the training while preserving the invariance of optimal policies. Therefore, we utilize the reshaped reward function based on the marginal utility our p gains over the prompt p' random baseline retrieves: $\mathcal{R}(p) = \mathcal{R}(p, \mathcal{Q}_{\tau_i}^{\text{query}}) - \mathcal{R}(p', \mathcal{Q}_{\tau_i}^{\text{query}})$.

3 Results

To evaluate the performance of the proposed MVPS method, we conduct experiments on three medical image modalities (dermatology, ophthalmology, and radiology) and four tasks (skin lesion segmentation, lung lesion segmentation, optical disc segmentation, and vessel segmentation) over 11 datasets.

Datasets. For *dermatology image segmentation* task: we use the HAM10k [25] dataset for meta-training and validation; we test on PH2 [17] and ISIC2016 [8] datasets to study the relationship between prompt retrieval effectiveness and **distribution shifts** between source and target domains, including task/label shifts, population shifts, and technology shifts. For *ophthalmology image segmentation* task: we design two subtasks to explore prompt retrieval behavior under **contrasting task difficulties**. For optical disc segmentation, an easier task, we train on the Refuge [18] dataset and test on a combined set from PAPILA [13]+IDRID [19]. For the more challenging vessel segmentation task, we use the FIVES [11] dataset as the source, a combined DRIVE [23], CHASEDB [6], and STARE [9] datasets as the test sets. For *Chest X-ray COVID-19 identification* task: we use the CovidQUEX dataset as train set. Collected by the same group, the COVID-19 Radiography dataset [3] is used for meta-testing, examining a scenario where the source and target share the **same task** and minimal shifts.

Experiments and Comparisons. For prompt retriever, we use a transformer with an embedding size of 256, a max sequence length of 1101 (including 1000 support size, 100 query size, and 1 separator token), 8 heads, 8 encoder blocks, 4 decoder blocks, and a feedforward dimension of 2048. We sample 150000 meta-tasks for each modality and use a mix-up ratio of 0.1 for augmentation. We use an Adam optimizer with a learning rate of 1e-4 and batch size of 64 for 10 epochs. We use SegGPT [27] as the LVM backbone and a LAION-2B ViT model for embedding extraction. Experiments are done on 1 AWS-g5.12xlarge instance. We compare against random selection, TopK/Unsupervised approaches and the SupPR [32] as data-centric prompt retrieval approaches. We also compare with model-centric low-rank adaptation (LoRA) methods [10, 33] for LVM adaptation to medical domain, where we finetune 14% of total params on the train

set with full labels. To avoid randomness, we run experiments of 30 runs and report average scores as results. Sampling at a step size of $k = 1$, we test the regular MVPS frozen during test time and MVPS with test time adaptation (TTA) receiving policy gradient updates during test time, which assumes initial access to a dynamically expanding tiny labeled evaluation set as more prompt images get labeled. We also show that MVPS is a flexible module that can be combined with model-centric approaches like LoRA. Finally, we approximate a performance upper bound using supervised SOTAs [29].

Findings. Tested at prompt length $k = 2, 4, 8, 16, 32$ (for retinal images, we stop at 16 due to small test size), MVPS offers consistent ICL enhancements under various source-target settings, including the same task, different task difficulties, and different distribution shifts. Qualitatively, we found that better prompts found by the retriever help the model better recognize noisy patterns adjacent to lesions and not just segment based on object contrasts, as shown in Fig 3. Quantitatively, MVPS and MVPS+TTA average a 4.08% and 4.66% gain over TopK and a 7.40% and 7.97% gain over random selection. For the **same task** and minimal distribution shifts setting in X-rays, we observe a 3.6% performance gain of the frozen LVM over TopK but diminishing returns as the number of k increases. When we combine MVPS with LoRA, it reaches a 95.1% comparable to supervised SOTA at $k = 32$.

For **contrasting task difficulties** in Ophthalmology experiments, we found that the harder task (Vessel) has less (4.87%) but still significant average gain compared to the easier task (Optical Disc) (6.725%), similar to those findings in NLP prompting [31] and suggesting that prompt insights generalize across tasks. As shown in Table 1, for different **distribution shifts** in Dermatology, we found that MVPS shows consistent gains despite population shifts (ISIC) and technology shifts (PH2), while stronger label shifts in ISIC dataset may be the cause of less average gain than PH2 dataset. As an **ablation**, we found that our test time adaptation indeed helps improve performance further upon MVPS, yet we have smaller gains when the prompt pool is small in PH2 or retinal datasets. **Efficiency** With frozen LVMs, MVPS only needs to train the prompt retriever with trainable parameters of around 22M. For a comparison, parameter-efficient fine-tuning methods like LoRA have 52M trainable parameters, while fully fine-tuning LVMs such as medicalSAM [29] will require 260M parameters. SupPR requires contrastive pretraining of a vision transformer as a feature extractor (315 million parameters). It is also a flexible module that isn’t exclusive to other model-centric methods like LoRA, with our results showing a consistent gain of 2.93% improvement when combined. We also tested MVPS on other in-context learning backbones such as Universeg [1] on the PH2 dataset and still got a gain of 3.70% over TopK and our combined approach 4.92% over LoRA.

4 Conclusion

To enhance ICL efficiency when applying LVMs to new medical imaging domains, we propose MVPS, a Meta-driven Active Prompt Selection framework. We meta-

Method	Type	K	PH2	ISIC	Xray	K	Optical	Vessel
Random	Data	4	79.8	63.1	77.8	4	46.1	41
		8	80.1	67.9	82.1	8	64.4	59
		32	81.5	70.5	87.2	16	73	62.3
TopK [24, 30, 32]	Data	4	84.0	67.2	79.0	4	55	44.0
		8	84.2	72.3	84.3	8	65.5	61.0
		32	85.8	76.5	88.3	16	73.5	64.5
SupPR [32]	Data	4	88.17	72.69	81.95	4	59.77	47.97
		8	88.31	76.18	85.88	8	73.43	70.65
		32	89.28	77.20	88.10	16	79.22	72.53
MVPS	Data	4	88.2 _{↑4.2}	72.3 _{↑5.1}	82.6 _{↑3.6}	4	59.9 _{↑4.9}	48.6 _{↑4.6}
		8	88.5 _{↑4.3}	76.2 _{↑3.6}	86.2 _{↑1.9}	8	73.6 _{↑8.1}	71.0 _{↑10.0}
		32	89.5 _{↑3.7}	77.9 _{↑1.4}	89.1 _{↑0.8}	16	79.9 _{↑6.4}	72.6 _{↑7.1}
LoRA [33, 10]	Model	4	85.2	75.6	83.8	4	73.4	61.2
		8	86.3	78.2	88.8	8	79.3	72.5
		32	91.5	82.3	93.3	16	86.9	74.6
LoRA+MVPS	Combined	4	89.9 _{↑4.7}	78.5 _{↑2.9}	85.9 _{↑2.1}	4	76.5 _{↑3.1}	68.6 _{↑7.4}
		8	90.3 _{↑4.0}	79.4 _{↑1.2}	89.8 _{↑1.0}	8	81.7 _{↑2.4}	75.3 _{↑2.8}
		32	93.3 _{↑0.8}	92.3 _{↑2.8}	95.1 _{↑1.8}	16	89.7 _{↑2.8}	75.7 _{↑1.1}
Supervised [29]	-	-	96.4	94.8	98.1	-	96.6	86.8

Table 1: DICE Score Over Different Datasets and K. Full results in Appendix. DICE was used as both our scoring function and performance metric, for the convenience of showing the score of supervised methods as an upper bound for direct comparison. We have also tested the mIOU as scoring function and observed similar patterns.

Method	Type	K	PH2	ISIC	Xray	K	Optical	Vessel
MVPS	Data	4	88.2	72.3	82.6	4	59.9	48.6
		8	88.5	76.2	86.2	8	73.6	71.0
		32	89.5	77.9	89.1	16	79.9	72.6
MVPS+TTA	Data	4	88.2	72.8	85.4	4	59.9	48.6
		8	88.5	76.2	88.2	8	73.6	71.0
		32	90.9	78.3	89.5	16	81.1	73.2

Table 2: Ablation: Does Test Time Adaptation Help?

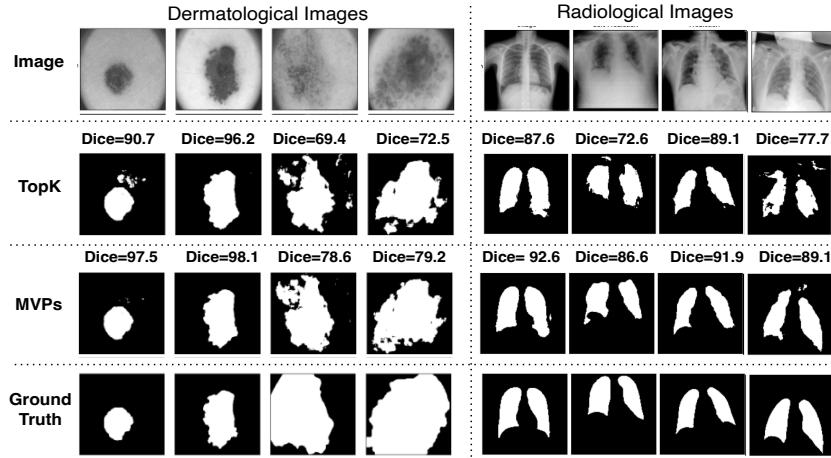


Fig. 3: Segmentation Results from MVPS vs TopK Prompting.

train a transformer-based prompt retriever optimized through reshaped reward policy gradients. MVPS offers a data-centric and finetuning-free enhancement to ICL, leading to efficiencies in training and label acquisition processes. MVPS’s flexibility allows for integration with different LVM backbones and model-centric approaches.

Acknowledgement. The authors acknowledge support from Michigan Institute for Computational Discovery and Engineering (MICDE) Catalyst Grant, and Michigan Institute for Data Science (MIDAS) PODS Grant.

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

References

1. Butoi, V.I., Ortiz, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Universalseg: Universal medical image segmentation. arXiv preprint arXiv:2304.06131 (2023)
2. Chen, Y., Huang, W., Zhou, S., Chen, Q., Xiong, Z.: Self-supervised neuron segmentation with multi-agent reinforcement learning. arXiv preprint arXiv:2310.04148 (2023)
3. Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., et al.: Can ai help in screening viral and covid-19 pneumonia? *Ieee Access* **8**, 132665–132676 (2020)
4. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
5. Deng, R., Cui, C., Liu, Q., Yao, T., Remedios, L.W., Bao, S., Landman, B.A., Wheless, L.E., Coburn, L.A., Wilson, K.T., et al.: Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:2304.04155 (2023)
6. Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A.R., Owen, C.G., Barman, S.A.: An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering* **59**(9), 2538–2548 (2012)
7. Geng, X., Liu, H., Lee, L., Schuurmans, D., Levine, S., Abbeel, P.: Multi-modal masked autoencoders learn transferable representations. arXiv preprint arXiv:2205.14204 (2022)
8. Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1605.01397 (2016)
9. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging* **19**(3), 203–210 (2000)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
11. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data* **9**(1), 475 (2022)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
13. Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., Sancho-Gómez, J.L.: Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data* **9**(1), 291 (2022)
14. Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P.: Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. arXiv preprint arXiv:2104.08786 (2021)

15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. *Nature Communications* **15**(1), 654 (2024)
16. Mavromatis, C., Srinivasan, B., Shen, Z., Zhang, J., Rangwala, H., Faloutsos, C., Karypis, G.: Which examples to annotate for in-context learning? towards effective and efficient selection. *arXiv preprint arXiv:2310.20046* (2023)
17. Mendonça, T., Celebi, M., Mendonca, T., Marques, J.: Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy image analysis* (2015)
18. Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis* **59**, 101570 (2020)
19. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabudhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data* **3**(3), 25 (2018)
20. Rauch, L., Aßenmacher, M., Huseljic, D., Wirth, M., Bischl, B., Sick, B.: Activeglae: A benchmark for deep active learning with transformers. *arXiv preprint arXiv:2306.10087* (2023)
21. Rubin, O., Herzig, J., Berant, J.: Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633* (2021)
22. Smit, A., Vrabac, D., He, Y., Ng, A.Y., Beam, A.L., Rajpurkar, P.: Medselect: Selective labeling for medical image classification combining meta-learning with deep reinforcement learning. *arXiv preprint arXiv:2103.14339* (2021)
23. Staal, J., Abramoff, M.D., Niemeijer, M., Viergever, M.A., Van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging* **23**(4), 501–509 (2004)
24. Sun, Y., Chen, Q., Wang, J., Wang, J., Li, Z.: Exploring effective factors for improving visual in-context learning. *arXiv preprint arXiv:2304.04748* (2023)
25. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* **5**(1), 1–9 (2018)
26. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6830–6839 (2023)
27. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1130–1140 (2023)
28. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* **8**, 229–256 (1992)
29. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620* (2023)
30. Zhang, J., Wang, B., Li, L., Nakashima, Y., Nagahara, H.: Instruct me more! random prompting for visual in-context learning. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 2597–2606 (2024)
31. Zhang, Y., Feng, S., Tan, C.: Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486* (2022)
32. Zhang, Y., Zhou, K., Liu, Z.: What makes good examples for visual in-context learning? *arXiv preprint arXiv:2301.13670* (2023)
33. Zhu, Y., Shen, Z., Zhao, Z., Wang, S., Wang, X., Zhao, X., Shen, D., Wang, Q.: Melo: Low-rank adaptation is better than fine-tuning for medical image diagnosis. *arXiv preprint arXiv:2311.08236* (2023)