Retriever Encoder Selection Matters for In-Context Learning-based Medical Segmentation

Fan Wang[1], **Zhongyi Han**[1*], **Yongshun Gong**[1], **Yilong Yin**[1*]

[1]School of Software, Shandong University, China
fanwangsail@gmail.com, zhongyi.han@sdu.edu.cn, ysgong@sdu.edu.cn, ylyin@sdu.edu.cn

Abstract

In-context learning-based medical segmentation (ICLM) enables foundation models to generalize to unseen cases without retraining. To enhance performance on test queries, existing methods typically follow a two-stage process: (1) using a retrieval encoder (RE) to map both queries and training samples into a shared feature space, and (2) retrieving and utilizing the top-k most similar training samples. While current methods fix the RE and focus on optimizing stage (2), we show that the choice of RE in stage (1) alone can account for over 70% of the performance variation, highlighting RE selection as a critical yet often overlooked factor in ICLM. In this paper, we conduct an analysis of the RE selection and make two main findings: (1) dynamically selecting the RE for each query outperforms selecting a fixed RE for the entire task; and (2) feature-space heuristics (*e.g.,* intra-class compactness and inter-class separability) fail to predict RE quality. To this end, we propose the *instance-adaptive retrieval encoder selection* (IRES) method that can select the optimal RE for each query based on output predictions. IRES is based on the intuition that a good RE retrieves relevant demonstrations, helping the ICL model generate more accurate and stable segmentation masks. Thus, we introduce the *shape stability score* (S^3), which evaluates the morphological stability of predicted masks under iterative erosion. Experiments show S^3 correlates strongly with true RE quality (Pearson > 0.8), serving as a reliable selection proxy. To reduce S^3's per-query cost, we propose *parallel prediction with reciprocal neighbor reuse* (P2R), which accelerates inference by parallelizing encoding and reusing encoder selections across reciprocal neighbors, avoiding redundant computation. Built on S^3 and P2R, IRES improves ICLM performance across FUNDUS, Brain MRI, and Chest X-ray datasets, with up to 10.6% gain on fundus segmentation.

Introduction

In-context learning (ICL) has emerged as a key mechanism in foundation models (Li et al. 2025) for medical image segmentation, such as UniverSeg (Butoi et al. 2023), ICL-SAM (Hu et al. 2024), and Tyche (Rakic et al. 2024). By conditioning on a few image–label demonstrations, ICL enables models to adapt to new tasks without retraining.

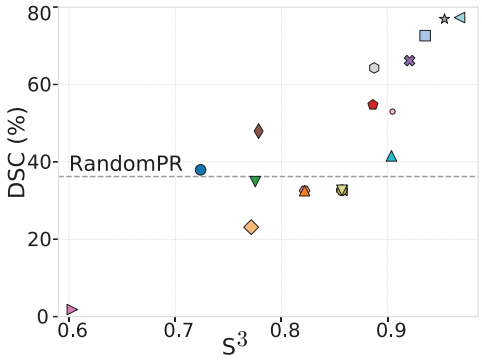


Figure 1: RE choice directly impacts ICL performance on optic disc segmentation: twenty REs (shapes in figure) yield $> 70\%$ DSC variation for the same query. Higher S^3 scores correlate with better performance.

This reduces the need for labeled data and improves generalization to domain shifts and diverse clinical environments (Zhang et al. 2022; Kim et al. 2023). Unlike natural-image tasks, medical task benefits more from ICL due to limited annotations and substantial domain variability.

To improve the performance of in-context learning-based medical segmentation (ICLM), selecting a few informative demonstrations from the training set for per test query is essential (Liu et al. 2024). Existing methods (Liu et al. 2024; Suo et al. 2024) fix a retriever encoder (RE)—a pretrained model that embeds queries and training samples into a shared feature space—and retrieve the top-k most similar training samples as ICL demonstrations. Assuming a general-purpose RE suffices, they focus on retrieval while ignoring the RE's impact. We show that RE choice—made before retrieval—strongly affects ICLM performance. Given the diversity of medical data across organs, modalities, and protocols, a single RE often fails to generalize. As shown in Figure 1, twenty different REs produce varied top-1 demonstrations for the same query, causing over 70% fluctuation in ICL performance—sometimes worse than random retrieval.

In this paper, we study, for the first time, the problem of automatically selecting the optimal RE for ICLM. This involves choosing the most suitable RE—from a predefined set of pretrained encoders—for the entire task or each test query to retrieve relevant demonstrations (see Figure 2). Our

*Corresponding authors.

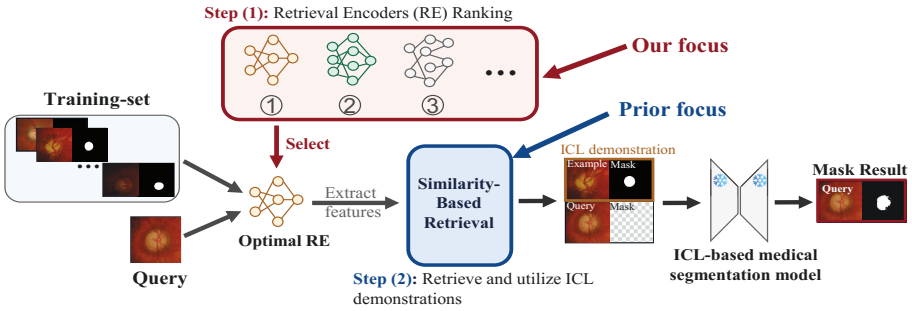


Figure 2: Unlike prior methods that focus on similarity-based retrieval with a fixed retrieval encoder (RE), this paper focuses on ranking and selecting the optimal RE for each input. This adaptive selection improves feature representation, enhances demonstration relevance, and leads to better in-context learning performance in segmentation tasks.

analysis yields two key findings. First, dynamically selecting the RE per query significantly outperforms using a fixed RE across all queries (see Figure 4), highlighting the impact of intra-domain variation. Second, naive strategies such as random choice or feature-based heuristics (*e.g.,* intra-class compactness, inter-class separability) fail to identify the optimal RE, showing weak correlation with segmentation performance (Pearson's $|\tau| < 0.2$; see Figure 3(c)). This is due to limited medical training data failing to capture both intra- and inter-class features effectively. These findings indicate that effective per-query RE selection is not just a minor detail, but essential for robust ICLM performance.

Motivated by the above insights, we propose *Instance-adaptive Retrieval Encoder Selection* (IRES), which can select the optimal RE for each query based on the output prediction. IRES builds on the observation that a good RE tends to retrieve top-k in-context demonstrations that are structurally similar to the query and provide meaningful contextual guidance. When used for prompting, these demonstrations help the model generate pseudo segmentation-masks with stable shapes. To quantify this property, we propose the Shape Stability Score (S^3), which assesses the morphological consistency of pseudo-masks under erosion, a common operation that gradually removes pixels from object boundaries. The intuition is that high-quality masks with coherent structures will retain their overall shape under erosion, while poorly formed or noisy masks will erode rapidly. Thus, a higher S^3 reflects better structural integrity and serves as an indicator of RE quality. Empirically, S^3 correlates well with segmentation performance (Pearson's $\tau > 0.8$; see Figure 6), making it a reliable per-query proxy for RE selection. However, evaluating S^3 on all REs per query is costly. We propose parallel prediction with reciprocal neighbor reuse, which enables parallel inference and reuses encoder selections among reciprocal neighbors—mutually similar queries—to reduce redundancy within local clusters.

Our contributions are summarized as follows: (1) We show that retrieval encoder selection is a critical yet overlooked factor in ICL-based medical segmentation, often dominating performance. (2) We propose the IRES framework that selects the optimal RE per query using a shape stability score and scales efficiently through parallel prediction with reciprocal neighbor reuse. (3) IRES enhances real-world segmentation performance across datasets.

Related Work

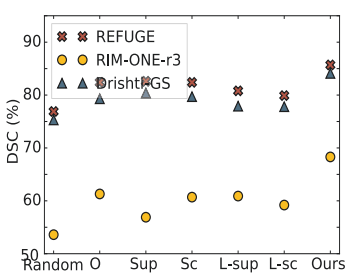
In-Context Learning-based Medical Segmentation

In-context learning (ICL) has shown strong potential in medical image segmentation, adapting across domains with only a few examples and no retraining—ideal for data scarcity and distribution shifts in clinical settings. Recent ICL-based medical segmentation (ICLM) models include Neuralizer (Czolbe and Dalca 2023) for single-shot inference, UniverSeg (Butoi et al. 2023) with paired prompts, ICL-MedSAM (Hu et al. 2024) combining SAM with prompt generation, and Tyche (Rakic et al. 2024) introducing prompt diversity via SetBlock. To improve ICL performance, existing methods typically retrieve and utilize demonstrations based on embedding similarity using a fixed retrieval encoder (RE). SupPR/UnsupPR (Zhang, Zhou, and Liu 2023), Prompt-SelF (Sun et al. 2025), SCS (Suo et al. 2024), InMeMo (Zhang et al. 2024a), and MVPS (Wu et al. 2024) refine this process through prompt fusion, filtering, or learned strategies. However, all these methods implicitly assume a universal RE, which is unrealistic in medical domains where high intra-domain variance often leads to suboptimal retrieval. We show that selecting task-specific REs before retrieval significantly boosts ICL performance.

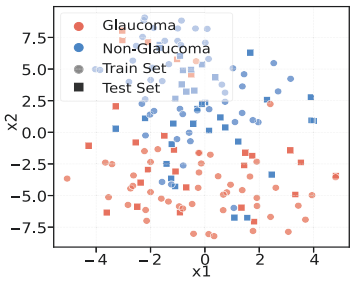
Pre-trained Model Selection

With the rise of large-scale pre-trained encoders, selecting a suitable model for downstream tasks without costly fine-tuning has become a key challenge. Existing model evaluation and selection methods are either probabilistic—using likelihood scores via linear probes (*e.g.,* LEEP (Nguyen et al. 2020), LogME (You et al. 2021), TransferScore (Yang et al. 2024), ProjectNorm (Yu et al. 2022))—or feature-based, evaluating class separability (*e.g.,* GBC (Pándy et al. 2022), H-score (Bao et al. 2019), Face (Ding et al. 2023)).

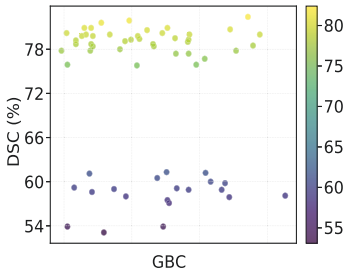
However, these methods are ill-suited for retrieval encoder selection in ICLM: probabilistic methods require output probabilities, which REs lack, and feature-based metrics are unreliable in medical ICLMs due to limited training data failing to capture stable class characteristics. To bridge this gap, we propose the first RE selection method for ICLM, directly selecting encoders that generate high-quality, instance-level demonstrations.



(a) Different methods



(b) TSNE on fine-tuned RN-50



(c) Naive encoder selection method

Figure 3: Limitations of fine-tuning retrieval encoders (RE) and naive encoder selection in ICL-MedSAM for fundus imaging. (a) Fine-tuning strategies—full supervision (Sup), supervised contrastive learning (SupCon) (Zhang, Zhou, and Liu 2023), and their high-layer variants (L-Sup, L-sc (= L-SupCon); fine-tuning only upper layers)—applied to a fixed RE (*e.g.,* RN50, with 'O' = no fine-tuning) on the training set fail to improve ICLM, as they do not retrieve more relevant demonstrations. (b) Supporting (a), the t-SNE of RN50 after Sup fine-tuning on RIM-ONE-r3 shows that while training samples are well separated, test queries remain entangled, confirming ineffective retrieval. (c) GBC scores (x-axis), a classic encoder selection metric, correlate weakly ($|\tau| = 0.13$) with true DSC (y-axis) across 20 encoders and 3 datasets in optic disk segmentation.

Understanding RE Selection in ICLM

To understand retriever encoder (RE) selection in ICLM, we first formalize the ICLM paradigm, then define RE selection within it, and finally address three central guiding questions.

In-Context Learning-based Medical segmentation (ICLM). Given a test query $x_q \in D_t$, ICLM retrieves a few relevant image–mask pairs from the training set D_s by mapping both the query and training images into a shared feature space using a fixed retriever encoder (RE) E, and selecting the top-k most similar samples: $\mathcal{P}(E, x_q, D_s) = \{(x_i, y_i)\}_{i=1}^{k}$. The segmentation model G then predicts the mask for x_q, conditioned on these retrieved demonstrations:

$$\hat{y}_q = G(x_q, \mathcal{P}(E, x_q, D_s)). \tag{1}$$

Retrieval Encoder Selection in ICLM. This paper shows that ICLM performance is highly sensitive to the choice of RE E (Figure 1), as it directly affects the quality of retrieved in-context demonstrations. To address this, we aim to rank encoders from a predefined zoo $\mathcal{Z} = \{E_1, \ldots, E_w\}$ and select the optimal one $E^* \in \mathcal{Z}$ for each input (Figure 2). RE selection can be conducted in two modes:

(1) *Instance-level:* For each query x_q, select E_q^* using a score $\mathcal{S}$ that is strongly correlated with segmentation performance:

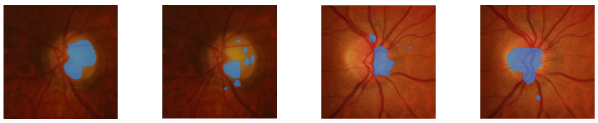
$$E_q^* = \arg\max_{E_z \in \mathcal{Z}} \mathcal{S}(E_z, x_q). \tag{2}$$

(2) *Task-level:* Selecting a single encoder E_t^* for the entire task. Given the training set D_s is labeled, we assess the suitability of each candidate encoder $E_z \in \mathcal{Z}$ by computing $\mathcal{S}$ over all training samples:

$$E_t^* = \arg\max_{E_z \in \mathcal{Z}} \frac{1}{|D_s|} \sum_{x_s \in D_s} \mathcal{S}(E_z, x_s). \tag{3}$$

Once the optimal encoder E^* is selected—either at the instance or task level—we use it to retrieve relevant in-context demonstrations and predict the segmentation mask for each test query $x_q \in D_t$:

$$\hat{y}_q = G(x_q, \mathcal{P}(E^*, x_q, D_s)). \tag{4}$$



(a) query A (RN-18 vs ViT-base) (b) query B (RN-18 vs ViT-base)

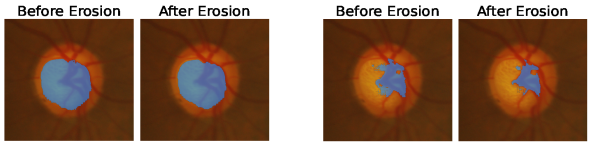
Figure 4: Selecting a fixed RE leads to inconsistent ICLM performance even for samples from the same domain: RN101 outperforms on query A (0.709 vs. 0.382), while VIT-base excels on query B (0.772 vs. 0.363), showing that no single RE guarantees optimal results.

Q1: Why does automatic RE selection matter in ICLM?

To evaluate the importance of RE selection in ICLM, we conduct two experiments: (1) Using ICL-MedSAM, we test 20 REs on optic cup segmentation. Different REs retrieve different Top-1 examples, causing over 70% variation in ICL performance (DSC; Figure 1), demonstrating the importance of RE choice. (2) One might ask whether fine-tuning a fixed RE on the training set (Zhang, Zhou, and Liu 2023) could match or even outperform RE selection. We compare our method with RN-50 fine-tuned using four strategies: fully supervised (Sup), supervised contrastive (SupCon) (Zhang, Zhou, and Liu 2023), and their high-layer variants (L-Sup, L-SupCon), where only upper layers are updated. As shown in Figure 3(a), our method—without any retraining—outperforms all fine-tuned variants by up to 5% DSC. This suggests that fine-tuning is suboptimal in medical domains due to sample scarcity and overfitting. Figure 3(b) supports this: fine-tuned REs tend to overfit—separating training data well but failing to generalize to test queries—leading to irrelevant retrievals and degraded ICL performance.

Q2: Should RE selection be per-query (Eq. (2)) or per-task (Eq. (3))?

We argue that RE selection should be done per query. First, intra-domain variation in medical data means queries from the same domain may rely on different features—some on local textures, others on global context. Figure 4 il-



(a) high-quality predicted mask (b) low-quality predicted mask

Figure 5: Higher-quality masks yield more stable structures under erosion. (a) High-quality mask (DSC 89.6%) shows minimal change after erosion (S^3 = 0.92). (b) Low-quality mask (DSC 48.7%) degrades significantly (S^3 = 0.67).

lustrates this: for a low-contrast retinal fold, the top-1 demonstration retrieved by ResNet-18 achieves a DSC of 0.709, much higher than ViT-base's 0.362; conversely, for bright capillary leakage, ViT-base reaches 0.772 compared to ResNet-18's 0.363. Second, our experiments show that selecting the optimal RE per query outperforms using a single best (oracle) RE for the entire dataset. As shown in Table 1, with UNIVERSEG on FUNDS-OD, per-query RE selection yields a 4.5% average improvement across three sub-datasets compared to oracle per-dataset selection.

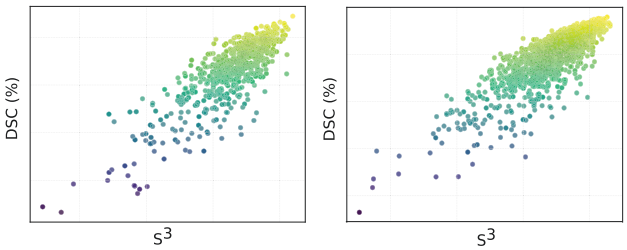
Q3: Why are naive RE selection methods ineffective?

To assess naive RE scoring, we evaluate GBC (Huang et al. 2022), a representative method that selects encoders based on feature space geometry—favoring high inter-class separability and low intra-class variance in the training set D_s. As shown in Figure 3(c), GBC exhibits weak correlation with actual ICLM performance on optic cup segmentation ($|\tau| = 0.13$) and often ranks suboptimal encoders highest (Tables 1, 2). Though effective on natural images, GBC fails in ICL-based medical segmentation due to: (1) limited medical data causing unstable class structure estimates; (2) task-level scoring that overlooks instance-level variation, crucial for demonstration retrieval in ICLM.

Instance-adaptive Retrieval Encoder Selection

Building on the insights above, we propose *Instance-adaptive Retrieval Encoder Selection (IRES)*—a framework that selects the optimal RE for each query to improve ICLM performance. The key intuition is simple: a good RE retrieves in-context demonstrations that are structurally aligned with the query, providing meaningful contextual cues. When used for prompting, such demonstrations help the model generate pseudo-masks with stable, coherent shapes. To capture this property, we introduce the *Shape Stability Score (S^3)*, which measures the robustness of predicted masks under morphological erosion. Intuitively, high-quality masks preserve their core structure under erosion, while noisy or irregular masks degrade quickly. However, computing S^3 across all REs for every query is computationally expensive. To mitigate this, we propose *Parallel Prediction with Reciprocal Neighbor Reuse (P2R)*—an efficient strategy that enables parallel inference across REs and reuses high-quality selections among reciprocal neighbors. The full IRES algorithm is presented in Algorithm 1, with detailed discussions of both S^3 and P2R provided below.

Shape Stability Score (S^3). Given a query image x_q, we



(a) REFUGE (τ = 0.83) (b) Drishti-GS (τ = 0.84)

Figure 6: S^3 is strongly correlated with the ground-truth DSC in optic cup segmentation. Each query yields 20 (S^3, DSC) pairs from different retrieval encoders, all of which are visualized. τ denotes Pearson correlation.

first retrieve its similar top-s demonstrations $\{x_r^j, y_r^j\}_{j=1}^s$ from the training set $\mathcal{D}_s$ using cosine similarity in the feature space of a RE E_z:

$$\{x_r^j, y_r^j\}_{j=1}^s = \text{Top}_s\left(\{\cos(E_z(x_q), E_z(x_i)) \mid x_i \in \mathcal{D}_s\}\right). \tag{5}$$

The pair $(x_q, \{x_r^j, y_r^j\}_{j=1}^s)$ is fed into G to generate a pseudo-segmentation mask:

$$\hat{y}_q = G(x_q, \{x_r^j, y_r^j\}_{j=1}^s). \tag{6}$$

To assess the structural quality of the predicted mask $\hat{y}_q$, we propose the Shape Stability Score (S^3), which captures how well a mask preserves its core structure under morphological erosion. This design is motivated by classical shape analysis: erosion removes pixels from object boundaries, with high-curvature or noisy regions vanishing first (Haralick, Sternberg, and Zhuang 1987; Hirata and Papakostas 2021). Hence, regular masks with smooth, coherent shapes remain largely intact through successive erosions, while irregular or fragmented masks degrade rapidly (see Figure 5).

Formally, let $\hat{y}_q^{(0)}$ denote the initial predicted mask, and define $\hat{y}_q^{(i)}$ as the mask after i iterations of binary erosion using a fixed structuring kernel K_s (shape element): $\hat{y}_q^{(i)} = \hat{y}_q^{(i-1)} \ominus K_s, \quad \text{where} \quad i = 1, 2, \ldots, t$. After t steps, we track the remaining foreground area $|\hat{y}_q^{(i)}|$. Stable masks exhibit minimal loss over iterations, indicating strong shape integrity. Based on this, we define S^3 as:

$$S^3(x_q, E_z) = \begin{cases} 0, & \text{if } |\hat{y}_q^{(0)}| \in \{0, A\} \\ 1 - \frac{1}{t}\sum_{i=1}^{t}\left(\frac{|\hat{y}_q^{(0)}| - |\hat{y}_q^{(i)}|}{|\hat{y}_q^{(0)}|}\right), & \text{otherwise} \end{cases} \tag{7}$$

Here, A denotes the total number of pixels in the image. When the zero-shot prediction $\hat{y}_q^{(0)}$ is either empty or fully covers the image (i.e., $|\hat{y}_q^{(0)}| \in \{0, A\}$), the segmentation model G is considered degenerate, and the stability score becomes unreliable. To ensure that S^3 reflects meaningful structural behavior, we define it as 0 in such cases. As a result, higher S^3 values indicate greater shape integrity and robustness, making it a reliable metric for evaluating pseudo-masks in ICL-based medical segmentation.

This erosion stability indicates that the pseudo-mask $\hat{y}_q$ is well-supported by the retrieved demonstrations $\{x_r^j, y_r^j\}_{j=1}^s$

Algorithm 1: Instance-adaptive Retrieval Encoder Selection

Input: ICL-based medical segmentation model G, training set D_s, test set D_t, encoder zoo $\mathcal{Z}$, cache Ω, parameters k, s, m, erosion steps t

for each query image $x_q \in D_t$ **do**
  if $x_q \in \Omega$ **then**
    continue {Skip if already processed before or via reciprocal neighbor reuse}
  end if
  Initialize candidate list $\mathcal{C} \leftarrow [\,]$
  Parallel for all retrieval encoders $E_z \in \mathcal{Z}$:
    Retrieve top-s demonstrations $\{x_r^j, y_r^j\}_{j=1}^s$ via Eq. (5), predict $\hat{y}_q = G(x_q, \{x_r^j, y_r^j\}_{j=1}^s)$ via Eq. (6)
    compute $\mathrm{S}^3(x_q, E_z)$ via Eq. (7), then append $(E_z, \mathrm{S}^3(x_q, E_z))$ to $\mathcal{C}$
  Select $E_q^* = \arg\max_{(E_z,\cdot)\in\mathcal{C}} \mathrm{S}^3(x_q, E_z)$ via Eq. (8)
  Predict final mask $\hat{y}_q = G(x_q, \mathcal{P}(E_q^*, x_q, D_s))$ based on Eq. (4)
  Identify reciprocal neighbors $\mathcal{R}(x_q)$ from D_t (Yang et al. 2021) in the feature space of E_q^*
  for each $x_j \in \mathcal{R}(x_q)$ and $x_j \notin \Omega$ **do**
    Predict and output: $\hat{y}_j = G(x_j, \mathcal{P}(E_q^*, x_j, D_s))$ based on Eq. (4), then add x_j to Ω
  end for
  Add x_q to Ω
end for

from encoder E_z. While prior work has used erosion to estimate mask uncertainty (Moses, Sammut, and Zrimec 2016; Wang and Wang 2021), we instead leverage S^3 to assess model-query compatibility in ICLM. As shown in Figure 6, S^3 correlates strongly with segmentation performance across datasets (*e.g.,* $\tau > 0.8$ on REFUGE and Drishti-GS for optic cup segmentation). Based on this, we rank encoders by S^3 and select the one with the highest score:

$$E_q^* = \arg\max_{E_z \in \mathcal{Z}} \mathrm{S}^3(x_q, E_z), \tag{8}$$

and use it to perform the final prediction with Eq. (4) for query x_q.

Parallel Prediction and Reciprocal Reuse (P2R). Reciprocal Reuse (P2R). Evaluating each query across all reference encoders (REs) in $\mathcal{Z}$ requires roughly $R \times I$ units of computation, where R denotes the number of REs and I represents the cost of a single forward pass through the segmentation model G. For N queries, the total computational demand increases proportionally to $N \times R \times I$. Although feature extraction and similarity search can be precomputed, and I remains constant during inference, the total cost still grows linearly with both N and R, limiting efficiency.

P2R mitigates this overhead through two components: *parallel prediction* and *reciprocal neighbor reuse (RR)*. The first leverages hardware parallelism to evaluate all R encoders concurrently. The second exploits local consistency in the feature space: queries that are *reciprocal neighbors*—i.e., each appears in the other's top-m nearest neighbors—tend to share semantic similarity (Yang et al. 2021; Wang et al. 2022; Zhang et al. 2024b) and often prefer the same RE. Based on this observation, RR directly assigns the optimal encoder E_q^* of a query x_q to its reciprocal neighbors: $\mathcal{R}(x_q) = \{x_j \mid x_j \in \text{top}_m(x_q),\ x_q \in \text{top}_m(x_j)\}$, thereby skipping the computation of selection scores (S^3) for these samples. This avoids redundant scoring and encoder evaluation, significantly accelerating inference. A cache Ω stores completed samples so that any $x_j \in \Omega$ is skipped in future steps. Moreover, RR enables encoder reuse within local clusters by propagating the selected encoder to its reciprocal neighbors, preventing redundant one-time usage.

In combination, P2R reduces the total inference workload to approximately $N' \times R' \times I$, where $N' < N$ denotes the subset of queries requiring full evaluation and R' reflects the effective cost amortized by parallel processing. On fundus image segmentation, P2R achieves more than a 60% reduction in inference time (Table 3) with less than a 2% drop in accuracy (Figure 7(a)), demonstrating strong efficiency with minimal performance degradation. The complete procedure is summarized in Algorithm 1.

Experiments

Dataset. Following (Hu et al. 2024), we evaluate IRES on three public medical segmentation tasks: (1) Fundus segmentation (FUNDUS-OD/OC): A two-class task segmenting the optic disc (OD) and optic cup (OC) using REFUGE (Orlando et al. 2020), RIM-ONE-r3 (Fumero et al. 2011), and Drishti-GS (Sivaswamy et al. 2015). (2) Brain tumor segmentation: A three-class task segmenting whole tumor regions from 2D axial FLAIR MRI slices in BraTS2020 (Bakas et al. 2018), including both high-grade (HGG) and low-grade gliomas (LGG). (3) Chest X-ray segmentation: A two-class task extracting lung masks from frontal chest X-rays (Jaeger et al. 2013; Candemir et al. 2013).

Retrieval encoder Zoo. We construct an encoder zoo comprising 20 widely used pre-trained retrieval encoders, spanning a range of architectures, training paradigms, and domain specializations. CNN-based encoders: ResNet-18/50/101 (Targ, Almeida, and Lyman 2016), VGG16 (Simonyan and Zisserman 2014), EfficientNet (Tan and Le 2019), MobileNet (Howard 2017). Transformer-based encoders: ViT-Base (Dosovitskiy et al. 2020), DeiT-Base (Touvron et al. 2021). Hybrid encoders: ConvNeXt (Liu et al. 2022). Contrastive learning-based encoder: DINOv2-Giant (Caron et al. 2021). Multi-modal vision encoders: EVA-Giant-Patch-CLIP (Fang et al. 2023), ViT-Large-Patch14-CLIP (Li et al. 2021, 2024). Medical imaging-specific encoders: MedSAM-ViT (Ma et al. 2024) (segmentation); RETFound-OCT and RETFound-CFP (Zhou et al. 2023) (fundus analysis). General medical vision-language encoders: PLIP (Huang et al. 2023), BiomedCLIP (Zhang et al. 2023), PathCLIP (Zheng et al. 2024), COACH (Lu et al. 2024), PubMedCLIP (Eslami, de Melo, and Meinel 2021).

Implementation Details. To validate our method, we compare it with three state-of-the-art ICL-based medical segmentation models (MSMs): UNIVERSEG (Butoi et al. 2023), ICL-SAM (Hu et al. 2024), and ICL-MedSAM (Hu et al. 2024). ICL-SAM and ICL-MedSAM automatically

		FUNDUS-OD				FUNDUS-OC				BrasTS2020		
MSMs	Method	REFUGE	RIM-ONE-r3	Drishti-GS	Avg	REFUGE	RIM-ONE-r3	Drishti-GS	Avg	HGG	LGG	Avg
UNIVERSEG	RandomPR	76.9	53.6	75.9	68.8	65.7	27.3	60.6	51.2	18.1	13.7	15.9
	Worst-S	77.4	53.9	75.9	69.1	64.7	28.4	56.1	49.7	15.3	8.5	11.9
	Avg-S	79.8	58.4	78.6	72.3	68.6	33.6	65.3	55.8	22.1	13.5	17.8
	GBC	80.0	58.1	78.5	72.2	70.9	35.3	67.4	57.9	24.3	14.2	19.3
	Best-S	82.4	61.3	80.9	74.9	**71.5**	**36.9**	**68.8**	**59.1**	27.9	**22.0**	**25.0**
	IRES	**85.7**	**68.3**	**84.1**	**79.4**	71.0	34.9	66.5	57.5	**32.3**	15.0	23.7
ICL-MedSAM	RandomPR	90.6	70.6	88.9	83.4	67.8	30.8	54.7	51.1	29.1	23.8	26.5
	Worst-S	90.5	70.3	88.9	83.2	65.9	33.9	54.1	51.3	23.8	20.0	21.7
	Avg-S	92.8	76.0	90.7	86.5	69.7	38.7	57.8	55.4	31.9	25.3	28.6
	GBC	92.5	77.3	90.4	86.7	72.7	40.3	57.9	57.0	34.7	**33.1**	33.9
	Best-S	94.7	79.0	91.5	88.4	**74.0**	42.7	60.5	59.1	38.2	**33.1**	35.7
	IRES	**95.0**	**85.6**	**93.7**	**91.4**	73.5	**48.9**	**66.6**	**63.0**	**43.4**	31.4	**37.4**
ICL-SAM	RandomPR	88.0	65.4	83.5	79.0	61.9	28.5	46.4	45.6	18.9	16.0	17.5
	Worst-S	90.4	61.0	85.6	79.0	62.2	31.3	45.2	46.2	13.9	10.9	12.4
	Avg-S	91.0	68.9	85.9	81.9	65.1	34.7	50.3	50.0	22.8	16.4	19.6
	GBC	89.8	71.3	86.2	82.4	69.4	34.3	51.4	51.7	24.1	14.3	19.2
	Best-S	91.4	73.3	85.1	83.3	**70.2**	36.9	52.7	53.3	29.2	26.4	27.8
	IRES	**91.5**	**77.1**	**89.5**	**86.0**	69.5	**43.8**	**57.4**	**56.9**	**35.1**	**29.7**	**32.4**

Table 1: Performance (DSC %) on FUNDUS and BrasTS2020. MSMs denote ICL-based medical segmentation models.

		Chext X-ray	FUNDUS (Distribution shift)	
MSMs	Method	LungS	FUNDUS-OD	FUNDUS-OC
UNIVERSEG	RandomPR	64.9	64.1	42.0
	Worst-S	64.6	59.4	37.2
	Avg-S	74.1	66.5	44.1
	GBC	65.0	67.5	44.3
	Best-S	76.1	70.1	**49.7**
	IRES	**80.3**	**72.6**	48.5
ICL-MedSAM	RandomPR	83.0	80.9	48.2
	Worst-S	82.6	77.3	40.7
	Avg-S	86.9	82.6	48.8
	GBC	82.7	83.2	47.3
	Best-S	88.1	85.9	56.6
	IRES	**92.0**	**88.4**	**58.0**
ICL-SAM	RandomPR	85.2	76.9	41.2
	Worst-S	84.4	72.9	36.2
	Avg-S	88.6	79.9	43.4
	GBC	85.7	80.1	43.3
	Best-S	89.8	**85.3**	51.3
	IRES	**92.1**	83.6	**52.0**

Table 2: Performance (DSC %) on Chest X-ray and FUNDUS (distribution shift).

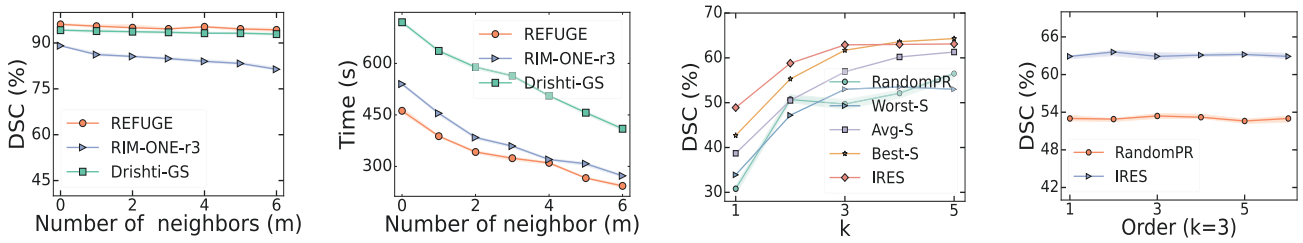
generate prompts from input images, eliminating the need for manual prompt design. Neuralizer (Czolbe and Dalca 2023) is excluded due to poor ICL performance on challenging Brain, and Tyche (Rakic et al. 2024) due to incomplete code. All models are evaluated without fine-tuning on NVIDIA Tesla V100 GPUs (40 GB), with results averaged over three random seeds. We set $s = 1$, $k = 1$ (in-context demonstrations), $m = 3$ (reciprocal neighbors), kernel size $K_s = 3$, and $t = 1$ for the S^3 module. Performance is reported using the Dice Similarity Coefficient (DSC).

Baselines. We compare IRES with two categories of baselines: (i) **Randomized retrieval**: (1) RandomPR: Randomly selects k in-context demonstrations from the training set for each test query. (ii) **RE selection methods**: (2) Worst-S: For each dataset, each of the 20 REs retrieves top-k demonstrations, which are then used to compute ICLM segmentation performance. The RE with the lowest performance is reported as the worst case. (3) Avg-S: Same as above, and reports the average ICLM performance across all 20 REs. (4) Best-S: Same as above, and reports the highest performance—an empirical upper bound. (5) GBC (Huang et al. 2022): Scores REs based on intra-class compactness and inter-class separability, following prior feature-level selection methods (Ding et al. 2023; Pándy et al. 2022; Bao et al. 2019). (6) IRES (Ours): Unlike prior methods that select a single encoder per dataset, IRES performs per-query encoder selection based on S^3 and P2R.

Note: All RE selection methods use the same top-k retrieval pipeline with cosine similarity, without post-retrieval strategies—ensuring a fair comparison focused on encoder selection, which is the focus of this work. Post-retrieval strategies (*e.g.,* Prompt-Self (Sun et al. 2025), SCS (Suo et al. 2024)) are beyond our scope and can be integrated in the future. Probability-based selection methods (Nguyen et al. 2020; Huang et al. 2022; Yu et al. 2022; Yang et al. 2024) are inapplicable as REs lack probabilistic outputs.

Main results

As shown in Tables 1 and 2, IRES consistently improves performance across three ICL-based medical segmentation models (MSMs) on three benchmarks, outperforming all baselines including RandomPR, Worst-S, Avg-S, and GBC. On the simpler FUNDUS-OD dataset (Table 1), IRES yields significant gains—for example, under UNIVERSEG, it improves DSC over RandomPR by 8.8%, 14.7%, and 8.2% on REFUGE, RIM-ONE-r3, and Drishti-GS, respectively. Similar improvements (>5%) hold across other MSMs. On the more challenging BraTS2020 dataset, IRES achieves a 14.9% average gain over RandomPR under ICL-SAM, showing strong adaptability to complex tasks, and consistently outperforms other baselines. For Chest X-ray segmentation (Table 2), IRES surpasses RandomPR by over 9% under both UNIVERSEG and ICL-MedSAM, and consistently outperforms all baselines, demonstrating IRES's generalization. Notably, IRES often outperforms the oracle



(a) Number of neighbors (b) Time efficiency (c) demonstration count (d) Order of demonstrations

Figure 7: Ablation study on the effects of reciprocal neighbor count, number of ICL demonstrations, and demonstration order on segmentation performance (DSC) and inference time.

MSMs	REFUGE			RIM-ONE-r3		
	t_o	t_{P2}	t_{P2R}	t_o	t_{P2}	t_{P2R}
UNIVERSEG	11.4m	7.0m	5.5m	13.2m	8.1m	5.8m
ICL-MedSAM	15.1m	7.7m	5.7m	16.7m	9m	6.4m
ICL-SAM	16m	7.7m	5.9m	15.3m	9.1m	7.2m

Table 3: Total selection and inference time on the full dataset using three ICL-based segmentation models (MSMs). P2R accelerates selection and inference via parallel prediction (PP) and reciprocal neighbor reuse (RR). t_o: original; t_{P2}: + PP; t_{P2R}: + PP + RR. Time in minutes (m).

Best-S—on FUNDUS-OD, it exceeds oracle performance by 3%–4.5% across MSMs—highlighting the strength of instance-level encoder selection in capturing fine-grained retrieval signals beyond task-level selection.

Further Analysis

Experiments on distribution shifts. We further assess IRES under domain shifts—a central challenge in medical segmentation. To this end, we design a cross-domain protocol on FUNDUS, where in-context demonstrations (training) and query samples (test) are drawn from different datasets. We define six shift scenarios: REFUGE→RIM-ONE-r3, REFUGE→Drishti-GS, RIM-ONE-r3→REFUGE, RIM-ONE-r3→Drishti-GS, Drishti-GS→REFUGE, and Drishti-GS→RIM-ONE-r3.

Table 2 reports average segmentation performance under shift, with three key observations: (1) RandomPR degrades notably under shift, dropping from 68.8% (Table 1) to 64.1% (FUNDUS-OD) and from 51.2% (Table 1) to 42.0% (FUNDUS-OC) under UNIVERSEG, indicating reduced utility of mismatched demonstrations. (2) Worst-S underperforms RandomPR in some cases, due to feature bias under domain shift impairing retrieval quality. (3) IRES consistently outperforms all baselines, improving performance by over 5% on average. It often exceeds even Best-S, showing that dynamic, instance-level encoder selection enhances robustness to distribution shifts by retrieving more relevant demonstrations.

Time analysis. We evaluate the efficiency of P2R in reducing the total selection and inference time of instance-adaptive encoder selection. Table 3 reports the ablation of our P2R on time in the FUNDUS-OD datasets (REFUGE and RIM-ONE-r3) across three ICL-based segmentation models (MSMs). For example, with ICL-MedSAM on REFUGE, encoder parallelism reduces runtime from 15.1 min (t_o) to 7.7 min (t_{P2}), and further to 5.7 min (t_{P2R}) with reciprocal neighbor reuse—highlighting the benefit of encoder sharing. This trend holds across datasets and MSMs, demonstrating P2R's efficiency and practical potential.

Effect of Reciprocal Neighbor Number. We assess the robustness of our RE selection IRES method under varying numbers k of ICL demonstrations. As shown in Figure 7(a), on simpler datasets (*e.g.,* REFUGE, Drishti-GS), increasing the number of reciprocal neighbors (m) has minimal effect on performance (within 1% of the baseline at m=0), indicating effective encoder reuse even with limited ICL samples. On the challenging RIM-ONE-r3, performance drops by up to 5% as m increases, likely because the shared encoder is suboptimal for distant neighbors. Figure 7(b) shows that larger m significantly reduces inference time (up to 60% at m=5) by cutting encoder selection overhead. We use m=3 by default to balance accuracy and efficiency.

Effect of Number and Order of In-Context Demonstrations. We examine the effects of demonstration number and order on optic cup segmentation using ICL-MedSAM on RIM-ONE-r3. As shown in Figure 7(c), performance improves with more demonstrations, with the largest gains from the first few; further additions yield diminishing returns, likely due to semantic redundancy. To assess order sensitivity, we test all six permutations of the top three demonstrations. Figure 7(d) shows performance varies by less than 0.5%, suggesting minimal sensitivity to order.

Conclusion

In this paper, we studied the overlooked practice of using a fixed, manually chosen retrieval encoder (RE) in in-context learning-based medical image segmentation (ICLM). Our analysis showed that segmentation performance varied by more than 70% across encoders for the same query, revealing the inefficiency of fixed REs. To address this, we formulated the retrieval encoder selection problem and demonstrated that instance-level selection led to substantial performance gains. We introduced Instance-adaptive Retrieval Encoder Selection (IRES), which dynamically selected the optimal RE for each query using a shape-based stability score and accelerated inference through reciprocal reuse and parallel prediction. Experiments on multiple datasets showed that IRES consistently improved ICLM performance.

Acknowledgments

This work is supported by the National Natural Science Foundation of China [U23A20389, 62176139] and by the Qilu Young Scholars Program of Shandong University.

References

Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R. T.; Berger, C.; Ha, S. M.; Rozycki, M.; et al. 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*.

Bao, Y.; Li, Y.; Huang, S.-L.; Zhang, L.; Zheng, L.; Zamir, A.; and Guibas, L. 2019. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE international conference on image processing (ICIP)*, 2309–2313. IEEE.

Butoi, V. I.; Ortiz, J. J. G.; Ma, T.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2023. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21438–21451.

Candemir, S.; Jaeger, S.; Palaniappan, K.; Musco, J. P.; Singh, R. K.; Xue, Z.; Karargyris, A.; Antani, S.; Thoma, G.; and McDonald, C. J. 2013. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2): 577–590.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Czolbe, S.; and Dalca, A. V. 2023. Neuralizer: General neuroimage analysis without re-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6217–6230.

Ding, Y.; Jiang, B.; Sheng, L.; Zheng, A.; and Liang, J. 2023. Unleashing the power of Neural Collapse for Transferability Estimation. *arXiv preprint arXiv:2310.05754*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Eslami, S.; de Melo, G.; and Meinel, C. 2021. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19358–19369.

Fumero, F.; Alayón, S.; Sanchez, J. L.; Sigut, J.; and Gonzalez-Hernandez, M. 2011. RIM-ONE: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, 1–6. IEEE.

Haralick, R. M.; Sternberg, S. R.; and Zhuang, X. 1987. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, (4): 532–550.

Hirata, N. S.; and Papakostas, G. A. 2021. On machine-learning morphological image operators. *Mathematics*, 9(16): 1854.

Howard, A. G. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, J.; Shang, Y.; Yang, Y.; Guo, X.; Peng, H.; and Ma, T. 2024. Icl-sam: Synergizing in-context learning model and sam in medical image segmentation. *Medical Imaging with Deep Learning*, 641–656.

Huang, L.-K.; Huang, J.; Rong, Y.; Yang, Q.; and Wei, Y. 2022. Frustratingly easy transferability estimation. In *International conference on machine learning*, 9201–9225. PMLR.

Huang, Z.; Bianchi, F.; Yuksekgonul, M.; Montine, T. J.; and Zou, J. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9): 2307–2316.

Jaeger, S.; Karargyris, A.; Candemir, S.; Folio, L.; Siegelman, J.; Callaghan, F.; Xue, Z.; Palaniappan, K.; Singh, R. K.; Antani, S.; et al. 2013. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2): 233–245.

Kim, S.; Lee, N.; Lee, J.; Hyun, D.; and Park, C. 2023. Heterogeneous graph learning for multi-modal medical data analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5141–5150.

Li, W.; Wang, Q.; Meng, X.; Wu, Z.; and Yin, Y. 2025. VT-FSL: Bridging Vision and Text with LLMs for Few-Shot Learning. *arXiv preprint arXiv:2509.25033*.

Li, W.; Wang, Q.; Zhao, P.; and Yin, Y. 2024. KNN Transformer with Pyramid Prompts for Few-Shot Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1082–1091.

Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Liu, R.; Li, M.; Zhao, S.; Chen, L.; Chang, X.; and Yao, L. 2024. In-context learning for zero-shot medical report generation. In *Proceedings of the 32nd ACM international conference on multimedia*, 8721–8730.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L. P.; Gerber, G.; et al. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3): 863–874.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Moses, D.; Sammut, C.; and Zrimec, T. 2016. Automatic segmentation and analysis of the main pulmonary artery on standard post-contrast CT studies using iterative erosion and dilation. *International journal of computer assisted radiology and surgery*, 11: 381–395.

Nguyen, C.; Hassner, T.; Seeger, M.; and Archambeau, C. 2020. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, 7294–7305. PMLR.

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. 2020. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59: 101570.

Pándy, M.; Agostinelli, A.; Uijlings, J.; Ferrari, V.; and Mensink, T. 2022. Transferability estimation using bhattacharyya class separability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9172–9182.

Rakic, M.; Wong, H. E.; Ortiz, J. J. G.; Cimini, B. A.; Guttag, J. V.; and Dalca, A. V. 2024. Tyche: Stochastic in-context learning for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11159–11173.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sivaswamy, J.; Krishnadas, S.; Chakravarty, A.; Joshi, G.; Tabish, A. S.; et al. 2015. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomedical Imaging Data Papers*, 2(1): 1004.

Sun, Y.; Chen, Q.; Wang, J.; Wang, J.; and Li, Z. 2025. Exploring effective factors for improving visual in-context learning. *IEEE Transactions on Image Processing*.

Suo, W.; Lai, L.; Sun, M.; Zhang, H.; Wang, P.; and Zhang, Y. 2024. Rethinking and improving visual prompt selection for in-context learning segmentation. In *European Conference on Computer Vision*, 18–35. Springer.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Targ, S.; Almeida, D.; and Lyman, K. 2016. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Wang, F.; Han, Z.; Gong, Y.; and Yin, Y. 2022. Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7151–7160.

Wang, Z.; and Wang, Z. 2021. Robust cell segmentation based on gradient detection, Gabor filtering and morphological erosion. *Biomedical Signal Processing and Control*, 65: 102390.

Wu, C.; Restrepo, D.; Shuai, Z.; Liu, Z.; and Shen, L. 2024. Efficient in-context medical segmentation with meta-driven visual prompt selection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 255–265. Springer.

Yang, J.; Qian, H.; Xu, Y.; Wang, K.; and Xie, L. 2024. Can We Evaluate Domain Adaptation Models Without Target-Domain Labels? In *ICLR*.

Yang, S.; Van de Weijer, J.; Herranz, L.; Jui, S.; et al. 2021. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in neural information processing systems*, 34: 29393–29405.

You, K.; Liu, Y.; Wang, J.; and Long, M. 2021. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, 12133–12143. PMLR.

Yu, Y.; Yang, Z.; Wei, A.; Ma, Y.; and Steinhardt, J. 2022. Predicting out-of-distribution error with the projection norm. In *International Conference on Machine Learning*, 25721–25746. PMLR.

Zhang, J.; Wang, B.; Li, L.; Nakashima, Y.; and Nagahara, H. 2024a. Instruct me more! random prompting for visual in-context learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2597–2606.

Zhang, S.; Xu, Y.; Usuyama, N.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; et al. 2023. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3): 6.

Zhang, T.; Fang, H.; Zhang, H.; Gao, J.; Lu, X.; Nie, X.; and Yin, Y. 2024b. Learning feature semantic matching for spatio-temporal video grounding. *IEEE Transactions on Multimedia*, 26: 9268–9279.

Zhang, Y.; Sheng, M.; Liu, X.; Wang, R.; Lin, W.; Ren, P.; Wang, X.; Zhao, E.; and Song, W. 2022. A heterogeneous multi-modal medical data fusion framework supporting hybrid data exploration. *Health Information Science and Systems*, 10(1): 22.

Zhang, Y.; Zhou, K.; and Liu, Z. 2023. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36: 17773–17794.

Zheng, S.; Cui, X.; Sun, Y.; Li, J.; Li, H.; Zhang, Y.; Chen, P.; Jing, X.; Ye, Z.; and Yang, L. 2024. Benchmarking pathCLIP for pathology image analysis. *Journal of Imaging Informatics in Medicine*, 1–17.

Zhou, Y.; Chia, M. A.; Wagner, S. K.; Ayhan, M. S.; Williamson, D. J.; Struyven, R. R.; Liu, T.; Xu, M.; Lozano, M. G.; Woodward-Court, P.; et al. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981): 156–163.