Show and Segment: Universal Medical Image Segmentation via In-Context Learning

Akhila Ravula
Texas Tech University
Texas, USA
R11909356

Phaninder Masapeta
Texas Tech University
Texas, USA
R11962048

Sriya Dhakal
Texas Tech University
Texas, USA
R118203

Scott Weeden
Texas Tech University
Texas, USA
R12041454

Zezheng Zhang
Texas Tech University
Texas, USA
R11913738

***Abstract*—Segmentation of medical images remains challenging due to the vast diversity of anatomical structures, imaging modalities, and segmentation tasks. Although deep learning has made significant advances, current approaches struggle to generalize as they require task-specific training or fine-tuning on unseen classes. We present Iris, a novel In-context Reference Image-Guided Segmentation framework that enables flexible adaptation to novel tasks through the use of reference examples without fine-tuning. At its core, Iris features a lightweight context task encoding module that distills task-specific information from reference context image-label pairs. This rich context-embedded information is used to guide the segmentation of target objects. By decoupling task encoding from inference, Iris supports diverse strategies from one-shot inference and context example ensemble to object-level context example retrieval and in-context tuning. Through comprehensive evaluation across twelve datasets, we demonstrate that Iris performs strongly compared to task-specific models on in-distribution tasks. On seven held-out datasets, Iris shows a superior generalization to out-of-distribution data and unseen classes. In addition, Iris's task encoding module can automatically discover anatomical relationships across datasets and modalities, offering insights into medical objects without explicit anatomical supervision.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Accurate segmentation of medical images is fundamental to modern clinical workflows, serving as a precursor for disease detection, volumetric analysis, surgical planning, and post-treatment monitoring. [1] High-quality organ and lesion segmentation enables precise diagnosis and targeted interventions in a wide range of medical conditions, including cancers, cardiovascular diseases, and neurological disorders [2]. However, achieving consistent and accurate segmentation in diverse patients and imaging modalities (e.g., CT, MRI) remains a long-standing challenge due to anatomical variability, imaging noise, acquisition protocols, and data scarcity. Conventional deep learning models, particularly Convolutional Neural Networks (CNNs), such as U-Net, V-Net, and their 3D variants, have demonstrated remarkable success in medical image segmentation when trained on sufficiently large and labeled datasets. [3] Despite their popularity, these models are inherently task-specific: each model is typically optimized for a particular dataset or organ of interest. As a result, adapting such a models to new tasks or imaging modalities often requires complete retraining or fine-tuning, which is computationally expensive and impractical in real-world clinical settings [3]. Furthermore, obtaining high-quality annotations for training is time-consuming and relies heavily on expert radiologists, making scalability a significant barrier. Recent research has explored the potential of foundation models and few-shot learning to generalize across tasks with minimal supervision. However, many of these approaches still fall short in handling volumetric 3D data or require cumbersome prompt engineering (e.g., SAM, MedSAM). The need for a unified, flexible, and generalizable segmentation model remains unmet.

To address these limitations, we explore Iris, a novel framework that introduces in-context learning (ICL) into the domain of universal medical image segmentation.[4] Inspired by advances in vision-language models, Iris formulates segmentation as a task-driven inference process, using reference image-label pairs to guide predictions without requiring retraining.[5] This reference-based paradigm fundamentally changes the inference workflow: the model is provided with examples of the target task during inference and generates predictions based on these contextual cues.

II. Related Works

Existing efforts to achieve "universal" medical image segmentation generally fall into three paradigms: (1) task-specific or universal multi-task models that learn fixed mappings but still require fine-tuning for unseen classes, (2) prompt-driven interactive models such as SAM and its medical variants that mandate human intervention, and (3) visual in-context learning (ICL) approaches that leverage reference examples but often suffer from efficiency limitations and single-class prediction. By reviewing these lines of work, we highlight the remaining gap—an automated, fine-tuning-free framework that is both accurate and computationally efficient.

A. Medical Universal Models.

Universal models for medical image segmentation seek to manage data variability across different tasks and modalities while acquiring broadly applicable feature representations. Initial approaches involved multi-dataset learning via label conditioning [12], implementing organ size constraints [62], and co-training with pseudo-labels [19]. More recent developments focus on advanced strategies for task encoding. DoDNet [55] was the first to utilize one-hot task vectors, which was

further developed in TransDoDNet [50] by incorporating transformer architectures. Notable advancements include CLIP-based models applying semantic encodings [30], integrating task-specific heads in MultiTalent [45], and utilizing modality priors in Hermes [13]. UniSeg [51] presented learnable task prompts, while MedUniseg [52] created a unified approach for handling both 2D and 3D images. Despite significant progress, these universal models still necessitate fine-tuning when encountering previously unseen classes. In contrast, Iris achieves segmentation of new classes with just a single reference image-label pair and does not require any model fine-tuning.

B. SAM-based Interactive Models.

The Segment Anything Model (SAM) is developed, which is one of the impressive contributions to interactive segmentation images, using a prompt-based model trained at scale and can be used to segment objects in natural images. Thanks to its success, adaptations specific to the special issues of medical imaging have been created in the form of medical variants. An example of this can be with MedSAM, which can use 1.5 million image-mask pairings, and SAM Med2D and SAM-Med3D can be able to extend and use 4.6 million 2D images or 2.2 thousand 3D volumetric datasets. The models are excellent where there are possibilities of user interaction, where segmentations can be adjusted back and forth. Their sensitivity to numerous prompts and hand-optimizations is, however, prohibitively time-consuming in automated, high-throughput settings, i.e., in high-throughput screening protocols or in the case of thousands of images in a research project. To overcome these challenges, Iris specifies segmentation tasks using context pairs (reference images and their matching masks) that can result in fully automatic processing at the expense of none of being able to provide adaptability. This architecture enables Iris to process a variety of medical imaging data, such as 3D volumetric data, with low user interaction, hence, more scalable in terms of clinical or research uses.

C. Visual In-context Learning.

In-context learning as introduced by GPT-3 [4] enables models to handle novel tasks through example-guided inference without a heavy retraining. In the vision community, Painter [48] and SegGPT [49] pioneered in-context segmentation through a mask image modeling framework. Alternative methods [33, 56, 44] explored SAM-based approaches through cross-image correspondence prompting, but their two-stage pipeline introduces redundant computation and heavily relies on SAM's capabilities, limiting their applicability to 3D medical images. Neuralizer [10] develop general tools on diverse neuroimaging tasks, like super-resolution denosing, etc. Recent works introduced specialized architectures for in-context segmentation [36]. For example, UniverSeg [5] is designed for in-context medical image segmentation, and Tyche [39] incorporated a stochastic inference. While these methods demonstrate promising capability on novel classes, they face two critical limitations. First, they show suboptimal performance compared to task-specific models on the training distribution. Second, they suffer from computational inefficiencies as they can only segment one anatomical class per forward pass, requiring multiple passes for multi-class segmentation. Meanwhile they must re-encode context examples for each query image even when using the same reference examples repeatedly. This becomes particularly problematic in high-throughput scenarios where multiple query images need to be processed. In contrast, Iris shows better performance and efficiency. An appealing design of our context task encoding module is to decouple task definition from inference, enabling the encoding of task from reference pairs into task tokens that can be efficiently reused across any number of query images, meanwhile multi-class segmentation can be done within a single forward pass.

The selection of appropriate context examples impacts the performance of in-context learning. Current methods [58] employ image-level retrieval strategies using global image embeddings to find better references. However, this approach faces significant challenges in medical image analysis where each image contains multiple classes including structures like organs, tissues, and lesions. Image-level retrieval inevitably averages features across all structures, leading to a suboptimal reference selection. To address this limitation, Iris introduces an object-level context selection mechanism that enables fine-grained matching of individual classes, focusing on more precise and class-specific reference selection compared to image-level approaches [58].

Moreover, most existing ICL frameworks exhibit an O(kmn) computational pattern—reference encoding is repeated per query and per class—resulting in prohibitive latency for large-scale 3D datasets. In contrast, Iris decouples task encoding from inference and performs multi-class segmentation in a single forward pass, effectively reducing the complexity to O(k+m).

III. Iris Architechture

Iris decouples task definition from segmentation inference. A lightweight Task Encoding Module compresses each support image–mask pair into reusable task embeddings, while a Mask Decoding Module consumes these embeddings to segment any query volume in a single forward pass. This design (i) avoids repeatedly encoding the same references, (ii) supports multi-class prediction at once

A. Task Encoding Module

Our task encoding module extracts task representations through two parallel streams to extract comprehensive task representations.The First One is Foreground feature encoding where Medical data volumes present unique challenges in feature extraction due to the presence of fine boundary details and anatomical structures spanning only a tiny portion of voxels. Direct feature pooling at downsampled resolution can lead to information loss or complete disappearance of these critical regions of interest (ROIs) and the other one is Contextual feature encoding where the above encoding

process extracted foreground features, but lacks important global context information. We encode these contextual information using learnable query tokens.To efficiently process high-resolution features while managing memory constraints, we employ strategy similar to sub-pixel convolution [43]. This approach permits a memory-efficient, high-resolution, feature-mask fusion.

Given one support pair (x_s, y_s), we first extract a downsampled feature map $F^s = E(x_s) \in \mathbb{R}^{C\times d\times h\times w}$ using a 3D encoder $E(\cdot)$. Two parallel branches then produce a *foreground descriptor* and a *contextual descriptor*. Their concatenation forms the task embedding for this class.

a) Foreground High-Resolution Branch: Low-resolution pooling tends to erase tiny targets. Therefore, we upsample F^s back to (near) input resolution, mask it with y_s, and apply global pooling:

$$T^f = \mathrm{Pool}\big(\mathrm{Up}(F^s) \odot y_s\big) \in \mathbb{R}^{1\times C}, \tag{1}$$

where $\odot$ denotes element-wise multiplication. T^f preserves fine boundary cues crucial for small structures.

b) Contextual Branch with Query Tokens: To encode global context efficiently, we adopt **PixelShuffle/Unshuffle** to trade channels for spatial resolution, concatenate the binary mask, and use a $1\times1\times1$ convolution for fusion. A set of learnable *query tokens* $Q \in \mathbb{R}^{m\times C}$ then interacts with these features through self- and cross-attention, yielding

$$T^c \in \mathbb{R}^{m\times C}. \tag{2}$$

c) Final Task Embedding: The final task embedding for one class is

$$T = [T^f; T^c] \in \mathbb{R}^{(m+1)\times C}. \tag{3}$$

d) Multi-class Handling: For K classes, we compute one T_k per class by reusing the shared encoder features F^s; only the (cheap) masking and pooling operations differ per class. All T_k are later concatenated for decoding.

B. Mask Decoding Module

The decoder D employs a query-based architecture [**?**] that efficiently handles both single and multi-class segmentation tasks. For a query image with features $\mathbf{F}_q \in \mathbb{R}^{C\times d\times h\times w}$, the task encoding module generates class-specific embeddings $\mathbf{T}^k \in \mathbb{R}^{(m+1)\times C}$ for each class k defined in reference image-label pairs. These embeddings are concatenated into a combined task representation $\mathbf{T} = [\mathbf{T}^1; \mathbf{T}^2; \ldots; \mathbf{T}^K] \in \mathbb{R}^{K(m+1)\times C}$, where K is the number of target classes and $K = 1$ for single-class segmentation. The bidirectional cross-attention mechanism processes this representation:

where $\mathbf{F}'_q$ and $\mathbf{T}'$ are the updated features. This mechanism enables effective information exchange between class-specific task guidance and query image features. T

Because task embeddings are computed once per support pair and cached, subsequent queries reuse them at negligible cost. Likewise, joint decoding avoids K independent predictions. Practically, when processing many queries with fixed supports, runtime scales almost linearly with the number of queries. We will quantify this in Section VII (Efficiency Study).

IV. Experimental Setup and Methodology

We will evaluate Iris across three key dimensions: in-distribution performance on trained tasks, out-of-distribution generalization to different domains, and adaptability to novel anatomical classes. Additional experiments analyze Iris's computational efficiency, inference strategies, and architectural design choices. Our training data comprises 12 public datasets [25, 3, 18, 21, 26, 22, 6, 41, 15, 35] spanning diverse body regions (head, chest, abdomen), modalities (CT, MRI, PET), and clinical targets (organs, tissues, lesions), split into 75,5,20 Percentages for train/validation/test. For out-of-distribution evaluation, we use 5 held-out datasets: ACDC [2], SegTHOR [24], and three MRI modalities from IVDM3Seg [16] to evaluate robustness against domain shift; MSD Pancreas (Tumor) [1] and Pelvic1K (Bone) [31] datasets are used for novel class adaptation. We randomly select 20 percentage samples from held-out sets for testing. Detailed dataset information is provided in supplementary materials.

A. Implementation Details

Iris employs a 3D UNet encoder trained from scratch using a one-shot learning strategy. We utilize the Lamb optimizer [**?**] with an exponential learning rate decay, configured with a base learning rate of 2×10^{-3} and a weight decay of 1×10^{-5}. Training spans 80,000 iterations with a batch size of 32 and includes 2,000 warm-up iterations. Data augmentation comprises random cropping, affine transformations, and intensity adjustments. Both training and inference operate on a volume size of $128\times128\times128$.

Training follows an *episodic* scheme that mimics test-time in-context segmentation. Each episode samples:

- a support set $S = \{(x_s^i, y_s^i)\}_{i=1}^k$, where x_s^i is a 3D volume and y_s^i its mask;
- one query volume x_q (and ground-truth y_q during training);
- a subset of classes for multi-class scenes; we randomly drop some classes to encourage class-wise independence.

B. Preliminary Results

We compare Iris against four representative categories of methods:

1) **Task-specific models:** nnU-Net (2D/3D variants).
2) **Universal multi-task models:** CLIP-driven, DoDNet/TransDoDNet, UniSeg, Multi-Talent, Hermes.
3) **SAM-based interactive/foundation models:** SAM, SAM-Med2D, SAM-Med3D[1].
4) **Visual in-context learning (ICL):** SegGPT, UniverSeg, Tyche-IS.

[1]Since these methods require user prompts, we simulate points/boxes using ground-truth masks to ensure a fair, fully automatic comparison.

Training was conducted on three datasets (AMOS CT, CHAOS, BCV) with a 75/5/20 train/validation/test split. Preliminary testing on AMOS CT for single-class liver segmentation achieved a mean Dice score of 80.12competitive with nnUNet's 83.18

V. Challenges

There are four related challenges that still exist in the development of Iris. i) Hard multi-class decoding. The existing cross-attention decoder is not optimised fully compatible with a large number of class-specific task embeddings to be processed simultaneously, resulting in unreliable masks across the queries in case multiple anatomical structures need to be segmented within a query volume. (ii) Computational efficiency. The extreme (128×128×128) resolution 3-D inputs monopolise GPU memory and run time, such that even a single A100 can be saturated, which makes large-scale or real-time clinical use impractical without gradient checkpointing or hybrid 2DD/33D crops. (iii) The diversity of datasets and generalisation. Training is confined to some CT/MRI sets (AMOS, CHAOS, BCV); no assuring generalisation to new and unseen modalities (e.g., PET), centres, or to rare classes (tumours, tiny vessels) can be made. The difference in kernel, slice thickness, and patient positioning between the centres of a cross-centre CT further worsens performance; histogram matching can only recover ĩpp Dice, so we are experimenting with feature-space registration losses and domain-oblivious image augmentations like elastic deformation or synthetically block-moving or deforming. Support selection ought to be efficient (iv). The naive cosine search of object-level retrieval grows linearly in a reference pool: 50,000 class embeddings already contribute 0.3s per query, which is 3 times longer than our query execution. Indexing with FAISS, a HNSW graph, or a learnable metric is under prototype because indexing all volumes with ¡ 1s cluster latency is clinically desirable. Last, one-shot robustness and flexibility are hard to balance; overfitting to specific references is harmful to queries with different behaviors in some situations-and a better reference-sampling timetable is under development.

VI. Current Work

A. Completed So Far

Since the mid-term checkpoint, we have finished all core components of the *task-encoding* module and verified their numerical stability under mixed-precision computation. In addition, a single-class version of the *mask-decoding* pipeline is fully operational: the bidirectional cross-attention block now runs end-to-end and already delivers **80.12 % Dice** on the AMOS liver subset, only three percentage points below the nnUNet baseline. Finally, we have wrapped the entire 3-D training pipeline—one-shot learning strategy, LAMB optimiser, and rich data augmentation (random cropping, affine transforms, and intensity jitter on 128^3 voxels)—into a reproducible script backed by Weights & Biases for automatic hyper-parameter sweeps and experiment tracking.

B. Ongoing Work

Our immediate focus is to extend the decoder from single-class to multi-class inference. The goal is to handle up to *15* anatomical structures in one forward pass while maintaining cross-attention efficiency. In parallel, we are running controlled experiments on *AMOS-CT*, *CHAOS*, and *BCV* to establish single-class baselines and diagnose multi-class scheduling issues. On the systems side, we are enabling activation checkpointing and mixed-precision inference to cap peak GPU memory below **7 GB** and to sustain a throughput of at least five 128^3 volumes per second on an NVIDIA A100.

VII. Conclusion

The Iris framework is a great contribution to universal image segmentation of medical images that can be used in clinical and research purposes molding a easy yet effective solution to the problem. With its Dice score of 80.12 on AMOS CT on single-class liver segmentation, it already shows the capability to compete with the task specific models such as imUNet, with the bonus attached of retraining-free flexibility. The completely built task encoding component successfully refines an example by constructing compact task representations, whereas the incompletely constructed mask decoding component has a potential draw in multi-class segmentation. Although problems such as dealing with multiplexing of anatomical classes, limitations associated with management of computational resources and possibilities of generalizing to different datasets, are of concern, it is being actively worked upon with specific optimizations being provided. By September 2025, Iris will provide a fully functional multi-class segmentation pipeline that will ensure high generalization model across modalities and anatomical structure, and therefore it will be ready to be used in high-throughput clinical workflow. Its capacity to do realistic learning in-context instead of retraining makes the framework stand as a transformative technology in medical imaging with the possibility of use in medical diagnostics, treatment planning, and research-based studies with thousands of participants.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, Jul. 2004.

[9] B. Li, Y. Liu, C. J. Occleshaw, B. R. Cowan, and A. A. Young, "In-line automated tracking for ventricular function with magnetic resonance imaging," *JACC, Cardiovascular Imag.*, vol. 3, no. 8, pp. 860–866, 2010

[10] P. V. Tran. (Apr. 2017). "*A fully convolutional neural network for cardiac segmentation in short-axis MRI*." [Online]. Available: https://arxiv.org/abs/1604.00494

[11] S. Petersen, "UK Biobank's cardiovascular magnetic resonance protocol," *J. Cardiovascular Magn. Reson.*, vol. 18, p. 8, Feb. 2016.

[12] C. Petitjean and J.-N. Dacher, "A review of segmentation methods in short axis cardiac MR images," *Med. Image Anal.*, vol. 15, no. 2, pp. 169–184, 2011

[13] S. C. Mitchell, J. G. Bosch, B. P. F. Lelieveldt, R. J. van der Geest, J. H. C. Reiber, and M. Sonka, "3-D active appearance models: Segmentation of cardiac mr and ultrasound images," *IEEE Trans. Med. Imag.*, vol. 21, no. 9, pp. 1167–1178, Sep. 2002.