Dual-Dynamic In-Context Learning for Medical Image Segmentation

Tiana Tianying Chen, Liangli Zhen, Yanyu Xu, Shishuai Hu, Huazhu Fu, and Limsoon Wong

***Abstract*— Visual in-context learning (ICL) enables medical image segmentation models to adapt to new tasks using only a small set of annotated support images, without additional model fine-tuning. This makes visual ICL promising for clinical deployment, yet a fundamental limitation remains. Current ICL models lack the intrinsic ability to adaptively integrate supports based on their relevance to the query. By either aggregating supports uniformly or relying on external retrieval, these methods limit query-aware support integration within the model itself. We introduce Dual-DSA (Dual-Dynamic Support Alignment), a reliability-aware visual ICL framework that dynamically integrates supports at the patch and support-set levels. At the patch level, DynPatch-Align selectively attends to relevant support regions for each query location, enabling fine-grained contextual alignment. At the support set level, evidential fusion integrates support-conditioned predictions based on uncertainty and consistency, and provides an Aggregate Evidential Confidence Score for reliability assessment. Together, these mechanisms enable intrinsic, query-adaptive support integration across both levels. Extensive experiments on 11 medical imaging datasets across endoscopy, breast ultrasound, and thyroid ultrasound show that Dual-DSA achieves strong full-coverage performance, obtaining the best Dice score on 8 out of 11 datasets and consistently outperforming its attention-based ICL backbone. Furthermore, the derived confidence scores correlate with segmentation quality and enable confidence-based rejection to improve retained-case performance, providing a practical mechanism for quality control in clinical workflows.**

***Index Terms*— dynamic support adaptation, evidential uncertainty modeling, medical image segmentation, visual in-context learning**

I. Introduction

Medical image segmentation is fundamental to medical image analysis, and has widespread applications spanning from disease diagnosis to treatment planning [1], [2]. However, deploying segmentation models in practice remains challenging, as models trained on source domains often generalize poorly to new clinical settings [3]–[6]. This generalization gap stems from domain shift across imaging equipment and protocols, heterogeneity in image characteristics across institutions, and anatomical variability across patients. These factors are all compounded by limited annotated data and privacy constraints that prevent centralizing data from multiple sites for training. While fine-tuning and domain adaptation on target data can improve performance [7], [8], these approaches demand large annotated datasets, substantial computational resources, and

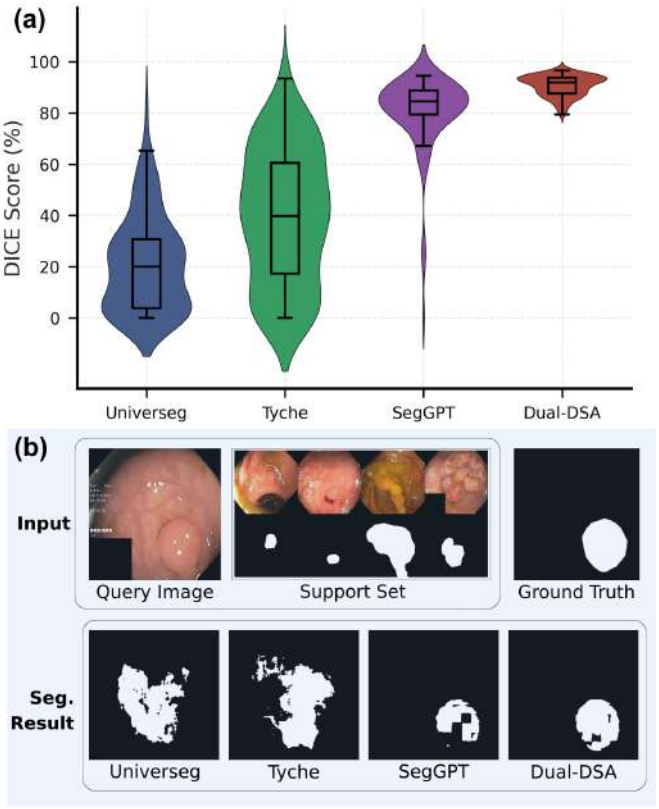

Fig. 1. Impact of support set sensitivity on segmentation performance. (a) Existing ICL models show significantly varied results with different support sets. We show the distribution of the Dice scores from a randomly selected query when using 100 different support sets for inference on UniverSeg, Tyche, SegGPT, and Dual-DSA. (b) Segmentation quality with the same support set. Dual-DSA robustly transfers target object from supports, despite differences in pose, scale, and position.

This work was supported by the National Research Foundation, Singapore, under the AI Singapore Programme (Award Numbers: AISG2-TC-2021-003 and AISG4-GC-2023-007-1B) and the National Medical Research Council, Singapore.

T. Chen and L. Wong are with the National University of Singapore, Singapore 117417, Singapore (e-mail: tianachen@u.nus.edu; limsoon@nus.edu.sg)

L. Zhen and H. Fu are with the Institute of Advanced Intelligence and Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632 (e-mail: zhenll@a-star.edu.sg; hzfu@ieee.org)

Y. Xu is with the Joint SDU-NTU Centre for Artificial Intelligence Research, Shandong University, China (e-mail: xu_yanyu@sdu.edu.cn)

S. Hu is with National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: sshu@mail.nwpu.edu.cn)

Corresponding authors: Liangli Zhen and Limsoon Wong

technical expertise, which are difficult to meet in most clinical environments.

Visual in-context learning (ICL) offers a promising alternative by enabling adaptation to new segmentation tasks without target-domain fine-tuning [9]–[11]. In this paradigm, users define the segmentation task through a support set comprising image-mask pairs that provide contextual guidance for segmenting a query image. Visual ICL extracts and utilizes support information by mapping support-query pairs to a shared feature space, where model-specific operations (*e.g.*, convolution or attention) establish correspondences between them. This approach holds considerable potential for clinical adoption, as it requires only a small number of annotated target-domain examples without retraining the model on the target data.

Existing ICL methods such as UniverSeg [9] have demonstrated promising results on medical image segmentation across diverse imaging modalities. Despite this progress, a fundamental limitation persists: support set sensitivity. Variations in pose, scale, position, and appearance of the target object across supports substantially affect segmentation quality, as the transferred contextual information is inherently dependent on support characteristics. As illustrated in Fig. 1, existing models exhibit high performance variability across different support sets (Fig. 1a) and produce inaccurate segmentation when support and query images differ substantially (Fig. 1b). Current visual ICL approaches attempt to mitigate support set sensitivity through two strategies. The first aggregates information from multiple supports, typically by averaging features from independent support-query pair mappings. This treats all supports as equally relevant, ignoring that some may align more closely with the query than others. The second strategy retrieves optimal supports from a large annotated database based on similarity metrics [12], [13]. This approach assumes the availability of such a database and that query images do not deviate substantially from its distribution. However, these assumptions rarely hold in clinical practice.

Critically, both strategies address support set sensitivity by modifying inputs rather than enhancing the model's intrinsic capability for adaptive support integration. In clinical scenarios where only a few annotated examples are available, visual ICL models must be able to leverage all supports flexibly, attending to them according to their relevance to each query. This raises a key question: *How can visual ICL models dynamically adapt from multiple supports for effective context transfer in medical image segmentation?*

To address this challenge, we propose **Dual-Dynamic Support Alignment (Dual-DSA)**, a novel ICL framework that dynamically aligns and integrates multiple supports at two complementary levels. Unlike existing methods that treat supports uniformly or rely on external retrieval, Dual-DSA intrinsically performs query-adaptive support integration within the model architecture. At the patch level, our DynPatch-Align module employs cross-attention to unify all support tokens within a shared attention space, enabling each query location to selectively attend to the most relevant regions across multiple supports. At the support set level, we use evidential fusion to integrate support-conditioned predictions according to their uncertainty and consistency, reducing the influence of unreliable support evidence. This evidential formulation also yields uncertainty estimates from which we derive an Aggregate Evidential Confidence Score that quantifies output reliability for clinical quality assurance. Together, these mechanisms realize dual-dynamic support alignment through spatially adaptive patch-level integration and reliability-aware support-set fusion, directly addressing support set sensitivity throughout the visual ICL pipeline.

Our main contributions are as follows:

- We propose Dual-DSA, a novel visual ICL framework that addresses support set sensitivity through intrinsic, query-adaptive support integration at two complementary levels, eliminating the need for external retrieval mechanisms or uniform support aggregation.
- We design a DynPatch-Align module that employs cross-attention to dynamically align query locations with relevant regions across multiple supports, enabling fine-grained contextual adaptation at the patch level. Additionally, we introduce support set-level alignment via evidential fusion, which fuses support-conditioned predictions based on their agreement and uncertainty, yielding an Aggregate Evidential Confidence Score for quantifying segmentation reliability in clinical workflows.
- We conduct extensive evaluation on 11 medical imaging datasets covering endoscopy, breast ultrasound and thyroid ultrasound imaging domains, demonstrating strong full-coverage performance and effective confidence-based rejection for reliability-aware segmentation.

II. Related Work

A. In-Context Medical Image Segmentation

The recent emergence of large vision foundation models, such as the Segment Anything Model (SAM) [14] and its medical variants [15], [16] has made zero-shot segmentation increasingly feasible. Users specify prompts, such as points, boxes, or text, to guide segmentation across tasks and domains.

While SAM and its variants typically require interactive prompts, visual ICL uses demonstration supports comprising a set of image-mask pairs, to automatically define the segmentation task. These visual ICL models are trained to learn the mapping between supports and query, and to transfer relevant features towards the query for effective in-context segmentation.

On natural images, transformer-based models such as VQ-GAN [10], Painter [11], and SegGPT [17] jointly encode support and query examples in a unified visual space by concatenating them into a grid. These models are trained with a masked-image-modeling objective to infer task relationships from contextual cues, enabling strong generalization across diverse visual segmentation tasks. In medical imaging, UniverSeg [9] and Tyche [18] adopt convolution-based architectures that learn support-query mappings via a cross-convolution mechanism, where support and query features are concatenated channel-wise and processed jointly.

Visual ICL is related to, but distinct from, conventional few-shot segmentation. Few-shot segmentation typically treats

support images and masks as a small labeled set for adapting to a new class or task, often through episodic training, prototype learning, metric learning, or support-conditioned adaptation modules. In contrast, visual ICL treats the support set as an in-context task specification that defines the desired segmentation behavior for the query. Rather than serving as task-specific adaptation data, the support set provides contextual guidance where visual ICL models infer support-query correspondences and transfer relevant visual information during inference.

Our work directly improves the robustness of the *support-to-query mapping mechanism* in in-context medical image segmentation by integrating information from multiple supports at both the query-representation and prediction levels.

B. Adaptive In-Context Learning

In many ICL settings, demonstration examples are provided as a set and treated as equally informative. Adaptive ICL addresses this limitation by adapting the in-context set to the query, either through support selection, weighting, or modulation. Prior work can be broadly categorized into retrieval-based [19]–[21] and weighting-based approaches [22]–[24].

Retrieval-based methods optimize the input space by selecting query-relevant examples from a large candidate corpus. These methods have been studied in visual ICL for both natural [12] and medical [13] domains. In contrast, weighting-based methods retain the support set, but modulate each example's contribution using similarity, uncertainty, or learned weighting functions as estimates.

Retrieval-based methods are less applicable in few-shot scenarios, where few annotated examples limit the available candidate pool [13]. Nonetheless, adaptive ICL for medical segmentation has primarily focused on retrieval mechanisms. Our work instead performs intrinsic query-adaptive support integration through spatially varying patch alignment and reliability-aware support-set fusion, without relying on large-scale retrieval.

C. Evidential Deep Learning for Uncertainty Estimation

Among deterministic approaches to uncertainty estimation, evidential deep learning (EDL) is an interpretable and efficient framework that models uncertainty as evidence supporting a distribution of predictions. This is achieved by reformulating the model's outputs as parameters of evidential distributions, such as the Dirichlet or Normal-Inverse-Gamma distributions, thereby avoiding multiple forward passes required by Monte Carlo sampling [25]–[27].

EDL has been applied to computer vision tasks such as open-set action recognition [28], semantic segmentation [29], and depth estimation [27]. In medical imaging, evidential formulations have also been explored for brain tumor segmentation [30] and whole-slide image analysis [31].

Our work leverages EDL for in-context medical segmentation, where uncertainty reflects both prediction reliability and the suitability of the support set for a given query.

III. Methodology

This section presents our method for improving the robustness of ICL for medical image segmentation. We begin by formulating the visual ICL segmentation task, then provide an overview of our framework, and finally describe the two core components: Dynamic Support Patch (DynSuppPatch) for query-adaptive patch-level alignment, and Dynamic Support Set (DynSuppSet) for reliability-aware support-set fusion.

A. Problem Formulation

We consider visual ICL for medical image segmentation, where the goal is to segment query images by leveraging annotated support examples without fine-tuning on the target domain.

Let $x_q \in \mathbb{R}^{H\times W\times 3}$ denote a query image for which we seek a segmentation mask $\hat{y}_q \in \{0,1\}^{H\times W}$. In this work, we focus on foreground (binary) segmentation, where each pixel is assigned either the target object foreground or the background. Accordingly, segmentation masks are defined as binary maps, and multi-class segmentation settings are not considered in this formulation. Visual ICL conditions prediction on a support set $\mathcal{S}$ comprising k image-mask pairs:

$$\mathcal{S} = \{(x_s^i, y_s^i)\}_{i=1}^{k}, \quad x_s^i \in \mathbb{R}^{H\times W\times 3}, \quad y_s^i \in \{0,1\}^{H\times W} \tag{1}$$

Here, subscripts q and s denote query and support quantities, respectively, and superscript i indexes support examples. Each support image x_s^i provides relevant context information about the target domain from which the query comes, with y_s^i providing its ground-truth mask. A visual ICL model f_θ produces predictions as:

$$\hat{y}_q = f_\theta(x_q, \mathcal{S}) \tag{2}$$

where parameters θ remain frozen during inference, and the adaptation occurs entirely through conditioning on $\mathcal{S}$.

In attention-based visual ICL, a common strategy is to condition the query on each support independently, then averaging the features. First, a single support and the query are combined and encoded jointly in a standard transformer encoder block:

$$\mathrm{Block}_{\mathrm{enc}}(z_q, z_s^i) = \mathrm{FFN}(\mathrm{LN}(z_q + \mathrm{SelfAttn}([z_q; z_s^i]))) \tag{3}$$

where z_q and z_s^i denote the query features and the i^{th} support features respectively, $[\cdot;\cdot]$ is the concatenation operation, $\mathrm{LN}(\cdot)$ denotes layer normalization, $\mathrm{SelfAttn}(\cdot,\cdot)$ is the self-attention operation, and $\mathrm{FFN}(\cdot)$ is a feed-forward network with residual connection. Next, all support-conditioned features are averaged to obtain the query output z_q':

$$z_q' = \frac{1}{k}\sum_{i=1}^{k} \mathrm{Block}_{\mathrm{enc}}(z_q, z_s^i) \tag{4}$$

After the last encoder block, z_q' is passed to a mask decoder to obtain the final prediction:

$$\hat{y}_q = \mathrm{Decoder}(z_q') \tag{5}$$

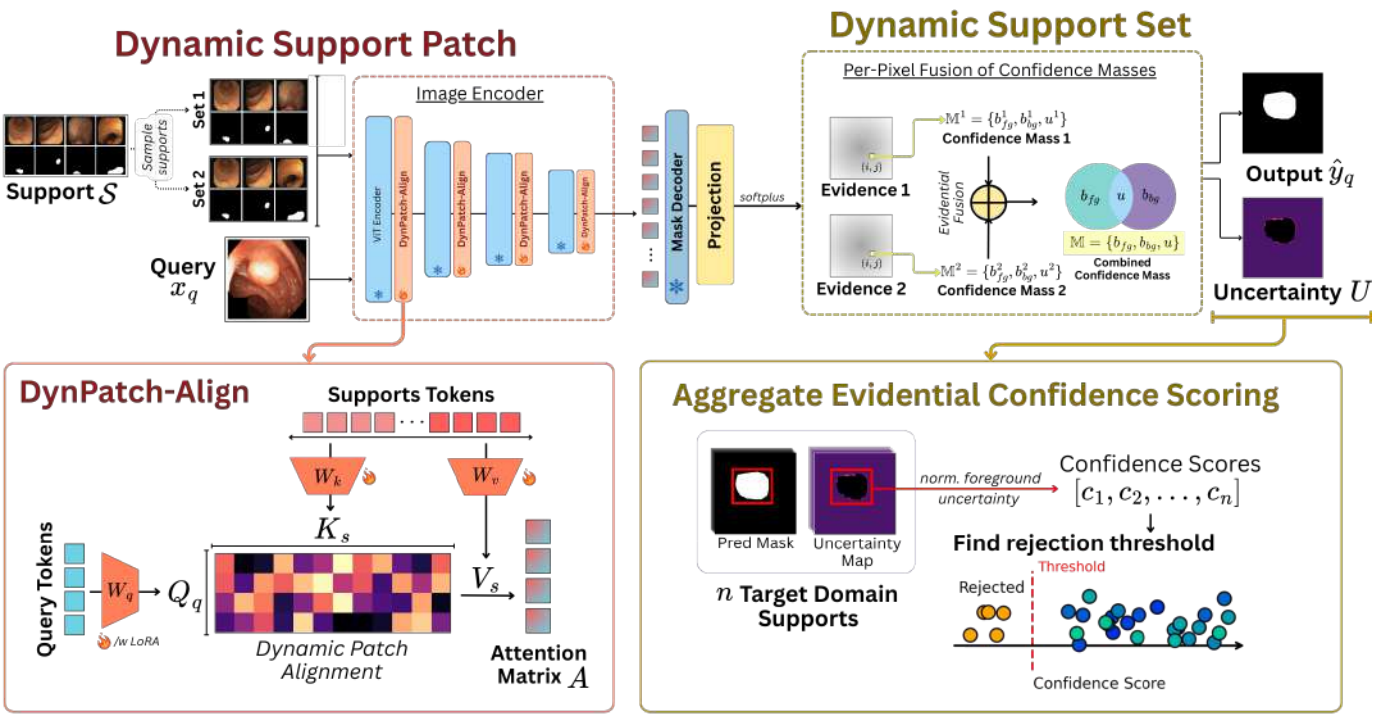

Fig. 2. Overall framework of our proposed method, comprising two core modules: Dynamic Support Patch (DynSuppPatch) and Dynamic Support Set (DynSuppSet). The DynPatch-Align module is integrated at selected encoder blocks and performs query-adaptive patch-level alignment across multiple supports. DynSuppSet uses evidential deep learning to perform reliability-aware fusion of two support-conditioned evidential predictions (Eq. 15). Aggregate Evidential Confidence Scoring converts the fused uncertainty map into an image-level confidence score.

B. Overview of Our Framework

Visual ICL models segment target objects by conditioning on support examples without task-specific fine-tuning. However, medical imaging presents substantial appearance variability due to acquisition protocols, patient anatomy, and pathology presentation. A single support rarely captures this diversity, necessitating multiple supports. Naively averaging features from multiple supports can dilute discriminative information when supports differ substantially. Furthermore, clinical deployment demands reliable uncertainty quantification to guide verification efforts. These challenges motivate adaptive support integration at two complementary stages, first within the query representation to identify relevant support regions, and second at the prediction level to account for inconsistencies in support-conditioned evidence.

As shown in Fig. 2, our framework comprises three integrated components: DynSuppPatch, DynSuppSet, and Aggregate Evidential Confidence Scoring. Given a query image x_q and a support pool $\mathcal{S}$, we sample two support subsets. Each support set and the query are processed through a shared vision transformer (ViT) encoder. Our **DynPatch-Align** module performs cross-attention between query tokens and concatenated support tokens, enabling dynamic patch-level alignment within the representation space. The enriched features pass through a mask decoder and projection layer to produce per-pixel class evidence, from which belief and uncertainty masses are derived. Processing two support sets produces support-conditioned evidential predictions, which undergo **per-pixel fusion** via Dempster's Rule of Combination, providing reliability-aware integration at the prediction level. Finally, **aggregate evidential confidence scoring** converts the fused uncertainty map into an image-level confidence score. Together, these components intrinsically integrate support information across representation and prediction spaces, enabling query-adaptive segmentation and reliability assessment within a unified framework.

C. Dynamic Support Patch Alignment

Unlike Eq. 4 where support interactions are independent and aggregated post hoc, our formulation enables joint interaction between each query token and *all support* tokens within a unified attention space. This enables the query representation $z_q \in \mathbb{R}^{L\times D}$ to attend directly to the entire support feature set Z_s:

$$z'_q = \mathrm{Block}(z_q, Z_s) \tag{6}$$

The Block function is implemented through our proposed **DynPatch-Align** module, which comprises cross-attention followed by a feed-forward network with residual connections:

$$\mathrm{Block}(z_q, Z_s) = \mathrm{FFN}(\mathrm{LN}(z_q + \mathrm{CrossAttn}(z_q, Z_s))) \tag{7}$$

where $\mathrm{CrossAttn}(\cdot,\cdot)$ is the cross-attention operation. Let L denote the number of patch tokens and D denote the token embedding dimension. Given k support feature representations $z_s^i \in \mathbb{R}^{L\times D}$, we first concatenate them along the token dimension to form a unified support representation:

$$z_s^{all} = [z_s^1; z_s^2; \ldots; z_s^k] \in \mathbb{R}^{kL\times D} \tag{8}$$

This design enables all support tokens to be projected into a shared key–value space, allowing each query token to evaluate support information jointly across all examples. Specifically, the query, key, and value matrices are defined as:

$$Q_q = \mathrm{W}_q z_q; \quad K_s = \mathrm{W}_k z_s^{all}; \quad V_s = \mathrm{W}_v z_s^{all} \tag{9}$$

where $W_q, W_k, W_v \in \mathbb{R}^{D\times D}$ are learnable projection matrices. The cross-attention output is computed as:

$$A = \mathrm{Softmax}\left(\frac{Q_q K_s^\top}{\sqrt{D}}\right) V_s \tag{10}$$

The softmax operation is applied along the support token dimension, normalizing each query token's attention weights over all kL concatenated support tokens. This joint representation enables simultaneous attention across supports, and avoids the need for independent pairwise processing followed by averaging. Consequently, the contribution of each support can vary across query locations, enabling fine-grained, query-conditioned alignment of relevant support regions. Importantly, concatenation here serves to enable a shared attention space. The influence of each support token is modulated through attention weights, allowing the model to emphasize informative regions while suppressing irrelevant or noisy support information. Thus, support relevance is resolved dynamically through joint cross-attention rather than imposed through fixed averaging, reducing sensitivity to uninformative support regions.

D. Dynamic Support Set Alignment

The second module employs evidential deep learning (EDL) to perform reliability-aware support integration at the prediction level. Following Dempster-Shafer theory under subjective logic, we derive belief and uncertainty from model evidence, and fuse predictions from multiple support sets to obtain a reliable segmentation. This complements DynPatch-Align by accounting for variations in the reliability of different support-conditioned predictions after decoding. Specifically, instead of directly predicting class probabilities, the model estimates non-negative evidence for each class, which parametrizes a Dirichlet distribution for confidence-aware inference.

1) Evidential Deep Learning: In segmentation, belief and uncertainty are defined pixel-wise. Since this work considers binary foreground segmentation, $C = 2$ denotes the background and foreground channels, and $n \in \{1, \ldots, C\}$ indexes the class channel. Given the query belief mass $\mathbf{b}_q \in \mathbb{R}^{H\times W\times C}$ and query uncertainty mass $u_q \in \mathbb{R}^{H\times W}$, a valid distribution satisfies:

$$\sum_{n=1}^{C} \mathbf{b}_q^n + u_q = 1 \tag{11}$$

The belief and uncertainty masses are derived from evidence $\mathbf{e}_q \in \mathbb{R}^{H\times W\times C}$ through the Dirichlet distribution. Let $\boldsymbol{\alpha}_q = \mathbf{e}_q + \mathbf{1}$ and $s_q = \sum_{n=1}^{C} \boldsymbol{\alpha}_q^n$ denote the concentration parameters and Dirichlet strength, respectively. Following subjective logic,

$$\mathbf{b}_q = \frac{\mathbf{e}_q}{s_q} = \frac{\boldsymbol{\alpha}_q - \mathbf{1}}{s_q}, \quad u_q = \frac{C}{s_q} \tag{12}$$

Here, $\mathbf{b}_q$ represents the normalized belief assigned to each class, while the uncertainty mass u_q captures uncommitted belief due to insufficient evidence. When the accumulated evidence is low (i.e., small Dirichlet strength s_q), uncertainty increases, allowing the model to express low confidence in ambiguous regions. To ensure valid evidence, we enforce non-negativity using a Softplus activation, which guarantees $\mathbf{e}_q \geq 0$ and stabilizes Dirichlet parameter estimation.

$$\begin{aligned} \mathbf{e}_q &= \mathrm{Softplus}(f_\theta(x_q, \mathcal{S})) \\ \boldsymbol{\alpha}_q &= \mathbf{e}_q + \mathbf{1} \end{aligned} \tag{13}$$

The final prediction is given by the mean of the Dirichlet distribution:

$$\hat{\mathbf{y}}_q = \frac{\boldsymbol{\alpha}_q}{s_q}, \quad \hat{y}_q = \arg\max_n(\hat{\mathbf{y}}_q^n) \tag{14}$$

For binary segmentation, this reduces to selecting the higher-probability channel between background and foreground at each pixel. This formulation enables principled fusion of predictions derived from multiple support sets via Dempster–Shafer theory, where both belief and uncertainty are explicitly considered during aggregation.

2) Evidential Fusion of Multiple Support Sets: To integrate evidence from multiple support sets, we apply Dempster's rule of combination from Dempster-Shafer theory. Each support subset produces a distinct support-conditioned evidential prediction, allowing uncertainty and agreement across predictions to inform their fusion. We employ two support sets for evidential fusion to maintain computational efficiency, while providing sufficient variation for reliability estimation. This approach balances robust uncertainty quantification with practical deployment considerations.

Dempster's rule conventionally combines distinct bodies of evidence, since dependence between sources may amplify shared information and lead to over-concentrated beliefs. To preserve sufficient contextual information within each support-conditioned prediction, we construct multiple support subsets with partial overlap rather than partitioning the available supports into smaller disjoint groups. This design relaxes the DST independence assumption, but ensures that each subset retains more contextual information while differing in composition, yielding partially dependent but distinct support-conditioned opinions. We use Dempster's rule to fuse these opinions according to their uncertainty and agreement, and further examine the corresponding disjoint-subset formulation in SectionV-D.3.

Given two support-conditioned opinions with mass assignments $\mathbb{M}_1 = \{\mathbf{b}_1, u_1\}$ and $\mathbb{M}_2 = \{\mathbf{b}_2, u_2\}$, the fused set $\mathbb{M} = \mathbb{M}_1 \oplus \mathbb{M}_2 = \{\mathbf{b}, u\}$ is computed as:

$$\begin{aligned} \mathbf{b}^n &= \frac{1}{1-M}\left(\mathbf{b}_1^n \mathbf{b}_2^n + \mathbf{b}_1^n u_2 + \mathbf{b}_2^n u_1\right) \\ u &= \frac{1}{1-M} u_1 u_2 \end{aligned} \tag{15}$$

where $M = \sum_{i\neq j} \mathbf{b}_1^i \mathbf{b}_2^j$ represents the conflicting mass, and $\oplus$ denotes Dempster's combination operator. This fusion reduces the influence of highly uncertain predictions. When one support-conditioned opinion exhibits high uncertainty, the

fused result primarily reflects the more confident opinion. Conversely, mutual agreement of confident opinions reinforces the combined belief.

E. Aggregate Evidential Confidence Scoring

To support clinical decision-making, we derive an image-level confidence score from the pixel-wise uncertainty map. Let $\mathcal{M}_q = \{(i,j) : \hat{y}_q(i,j) = 1\}$ denote the set of predicted positive pixels. The confidence score is defined as:

$$c_q = 1 - \frac{1}{|\mathcal{M}_q|} \sum_{(i,j)\in\mathcal{M}_q} u_q(i,j) \tag{16}$$

When no foreground pixels are predicted, the foreground-based confidence score is undefined. We therefore conservatively assign $c_q = 0$, treating empty predictions as unreliable and prioritizing them for review. This score represents one minus the mean uncertainty over predicted positive regions, providing an interpretable measure of prediction reliability.

During alignment to the target domain, we calibrate a rejection threshold using the support split. Each support sample is treated as a query while the remaining samples form the support set, yielding a distribution of confidence scores. In this work, we set the rejection threshold at the 10^{th} percentile of this distribution. During inference, predictions with confidence below this threshold are flagged for manual review, enabling clinicians to focus on verifying the least reliable cases.

IV. Experimental Study

A. Datasets

We evaluated our method on 11 datasets spanning three medical imaging domains: endoscopy, breast ultrasound and thyroid nodule ultrasound. These domains exhibit substantial variations in target pose, scale, size, location, and imaging conditions.

Polyps in endoscopy images show large pose, shape, and viewpoint variations due to non-rigid anatomy and continuous camera motion. Although breast and thyroid datasets are both ultrasound-based, they differ in anatomical structure, lesion morphology, and acquisition variability. Breast ultrasound images often contain highly heterogeneous backgrounds and ambiguous lesion boundaries. In contrast, thyroid nodules are smaller, increasing sensitivity to object scale and foreground sparsity. Together, these domains provide complementary settings for assessing the robustness of Dual-DSA under diverse and clinically meaningful conditions.

Kvasir-SEG [32] consists of 1,000 endoscopy images of polyps from the Vestre Viken Health Trust in Norway. **CVC-ClinicDB** [33] and **CVC-ColonDB** [34] include 612 and 380 annotated frames, respectively, from the Hospital Clinic of Barcelona. **ETIS-LaribPolypDB** [35] comprises 196 images from the ETIS laboratory. **BUSI** [36] contains 647 breast ultrasound images from Baheya Hospital, Egypt. **BUS-BRA** [37] comprises 1,875 breast ultrasound images from the National Institute of Cancer, Brazil. **BUS-UCLM** [38] includes 264 breast ultrasound images from Hospital General de Ciudad Real, Spain. **BUSI-WHU** [39] consists of 927 breast ultrasound images from Renmin Hospital of Wuhan University, China. **TNUI** [40], **TN3K** [41], and **DDTI** [42] include 1,378, 3,493, and 466 thyroid ultrasound images from Fujian Medical University Union Hospital, Zhujiang Hospital, and IDIME Imaging Center, respectively.

B. Implementation Details

1) Model: Our proposed framework is illustrated in Fig. 2, comprising DynSuppPatch at the encoder (orange box) and DynSuppSet at the outputs (yellow box).

The encoder is a vanilla ViT-Large with 24 transformer layers. To capture multi-level feature interactions, we use latent features from evenly spaced encoder depths at the 6^{th}, 12^{th}, 18^{th}, and 24^{th} transformer blocks as inputs to the decoder. DynPatch-Align is implemented at these depths by replacing the standard self-attention operation in Eq. (3) with the query-to-support cross-attention formulation in Eq. (7). Therefore, at these four blocks, query tokens are updated directly through cross-attention to the concatenated support tokens, and no separate query self-attention operation is performed before DynPatch-Align. The remaining transformer blocks retain their original self-attention operations, while query–support interactions are present throughout the network, applying DynPatch-Align at these selected depths enables joint alignment across multiple supports at different semantic levels. Specifically, earlier layers learn fine-grained spatial correspondence, whereas deeper layers capture higher-level semantic relationships. As such, the model integrates both local and global contextual cues, achieving a balance between spatial precision and semantic abstraction, rather than restricting to a single representation scale.

To learn evidence, we project the mask decoder's output logits using a linear evidential projection layer, followed by softplus rather than softmax, to obtain non-negative foreground and background evidence (Eq. (13)). This allows the evidential layer to estimate evidence from the final decoded representation, instead of converting the decoder output into class probabilities. The resulting evidence defines Dirichlet concentration parameters α. We obtain the subjective opinions (Eq. (12)) and combine the beliefs and evidence via Dempster's combination rule. From the combined opinions, we derive the predicted segmentation and the uncertainty map (Eqs. (12) and (14)).

2) Baselines: We compare Dual-DSA against three baseline categories: 1). Few-shot: **SENet** [43], **PANet** [44] 2). Interactive: **SAM-Med2D-B** (box), **SAM-Med2D-P** (point) [16], and 3). ICL models: **UniverSeg** [9], **Tyche** [18], and **SegGPT** [17]. SENet and PANet are representative few-shot segmentation models, which use information from support images and masks to generate conditioned weights or matching feature representations for segmenting the query. Although SAM-Med2D is originally designed for interactive segmentation, we adapt it into a few-shot setting by automatically deriving box and point prompts from support image-mask pairs, removing the need for human interaction. This ensures that SAM-Med2D operates under the same support-conditioned setting as ICL methods, enabling a fair comparison. Finally, we directly compare with three competitive ICL segmentation

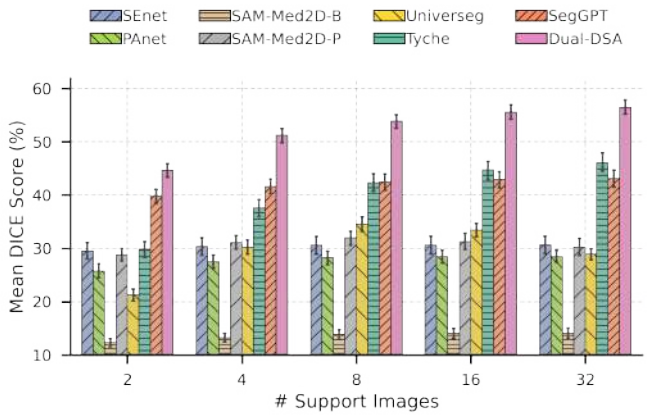

Fig. 3. Comparison of aggregated Dice (%) across support sizes. Dual-DSA achieves the highest overall performance, particularly at small support sizes (k=2,4). Results are aggregated across all datasets. Error bars denote 95% confidence intervals.

TABLE I
COMPARISON OF DUAL-DSA WITH BASELINES ON ENDOSCOPY DATASETS. DICE (%). ± STANDARD DEVIATION. DUAL-DSA REPORTS FULL-COVERAGE RESULTS, WHILE DUAL-DSA (REJ.) REPORTS POST-REJECTION RESULTS AFTER CONFIDENCE-BASED REJECTION.

Method	Kvasir-SEG	CVC-Clinic	CVC-Colon	ETIS
Few-shot				
SENet	25.38±0.021	23.57±0.079	6.37±0.021	17.30±0.037
PANet	51.85±0.024	34.16±0.040	22.82±0.035	12.34±0.037
Interactive				
SAM-Med2D-B	29.15±0.018	21.39±0.043	12.72±0.021	7.98±0.028
SAM-Med2D-P	44.01±0.027	35.05±0.065	24.62±0.037	14.36±0.045
In-context				
UniverSeg	31.01±0.012	23.89±0.034	15.74±0.026	9.38±0.024
Tyche	38.96±0.012	25.77±0.046	18.11±0.026	12.96±0.039
SegGPT	59.20±0.019	56.79±0.061	49.52±0.039	30.40±0.045
Dual-DSA	**65.98**±0.015	**59.29**±0.046	**52.29**±0.038	**35.08**±0.037
Dual-DSA (rej.)	70.45±0.015	64.24±0.050	57.77±0.026	41.88±0.052

models: UniverSeg, Tyche, and SegGPT, which also leverage support-query interactions to guide the segmentation task.

3) Training: We fine-tune models within each imaging domain using a cross-dataset adaptation setup, where one dataset is held out for evaluation and the remaining datasets are used for training. For Dual-DSA, we initialize the ViT-Large encoder and mask decoder using SegGPT pre-trained weights and perform offline medical-domain adaptation using low-rank adaptation (LoRA) [45] with rank $r = 2$ applied to the DynPatch-Align layers. All trainable models are fine-tuned with a support size of $k = 4$ for 20 epochs using the AdamW optimizer. For ICL models, we use LoRA with a learning rate of 1×10^{-5} to stably adapt pretrained encoders. For few-shot baselines without medical pretrained weights, we perform full fine-tuning, with a learning rate of 1×10^{-4}. SAM-Med2D is evaluated without fine-tuning, with inference-time prompts automatically derived from support masks.

Dual-DSA adapts pretrained support-query attention interactions from SegGPT to the medical domain, rather than learning these interactions from scratch. During training, support images and masks provide the in-context task specification, while supervision is applied only to the query prediction. Because DynPatch-Align updates the query representation by attending to support tokens, supervision on the query prediction also optimizes the support-query cross-attention layers. LoRA updates only a small subset of parameters while keeping most of the pretrained backbone fixed, focusing optimization on medical domain adaptation rather than relearning the full support-query mapping. This constrained adaptation helps stabilize training under randomly sampled support sets, and no noticeable training instability was observed across different random seeds. For the held-out target domain, no parameter updates are performed using the target-domain examples, and adaptation to each segmentation task is achieved solely through conditioning on the annotated support set.

To ensure fair comparison, we preserve each model's original input resolution and adaptation strategy whenever possible. For transformer-based models such as SegGPT and Dual-DSA, we use 224×224 inputs, which remain compatible with the pretrained transformer weights and allow training with multiple support images under identical settings. For evidential fusion, Dual-DSA constructs two support sets from $k = 4$ images, where each set contains three images sampled from $C(4, 3)$ combinations. The two sets may partially overlap, allowing each prediction to retain three support examples while being conditioned on a different support composition. Our methods are implemented with PyTorch on a single NVIDIA RTX 4090 GPU.

C. Evaluation

We evaluate our methods using the Dice score [46], a widely adopted metric in medical image segmentation. We perform five-fold cross-validation, where each fold is partitioned into a support split and a test split. The support split is used as a pool from which in-context supports are randomly selected for each query in the test split. For each query, we perform five independent predictions using different randomly selected support sets and report the mean Dice values across predictions and folds, which provides a robust estimate under stochastic support sampling. For Dual-DSA, we report full-coverage performance as the primary evaluation setting. We additionally apply Aggregate Evidential Confidence Scoring to identify samples below the rejection threshold and report post-rejection performance on the retained samples as a separate reliability analysis. To ensure a fair comparison, we use the same support images per query across all methods. Unless otherwise stated, all results are reported by evaluating with $k = 4$ supports.

V. RESULTS

A. Performance Comparison with the SOTA Methods

We compare Dual-DSA with existing few-shot, interactive, and ICL baselines on eleven public datasets across three imaging domains. Fig. 3 shows the aggregated full-coverage segmentation performance across all datasets for varying numbers of support images. Dual-DSA achieves strong performance with a small number of support images ($k = 2, 4$),

TABLE II
COMPARISON OF DUAL-DSA WITH BASELINES ON BREAST ULTRASOUND DATASETS. DICE (%).

Method	BUSI	BUS-BRA	BUS-WHU	BUS-UCLM
Few-shot				
SENet	**48.23**±0.035	34.11±0.017	55.90±0.015	37.84±0.039
PANet	28.99±0.026	27.36±0.009	27.44±0.016	35.39±0.039
Interactive				
SAM-Med2D-B	14.81±0.016	13.41±0.01	8.42±0.015	10.12±0.024
SAM-Med2D-P	32.49±0.023	43.49±0.014	44.95±0.019	34.43±0.034
In-context				
UniverSeg	30.87±0.018	42.95±0.013	44.39±0.020	38.66±0.029
Tyche	44.81±0.015	**56.02**±0.010	58.39±0.017	43.69±0.027
SegGPT	41.29±0.027	31.96±0.014	53.84±0.018	44.57±0.051
Dual-DSA	47.30±0.022	43.60±0.012	**59.98**±0.013	**50.30**±0.047
Dual-DSA (rej.)	53.29±0.022	51.39±0.014	66.18±0.015	57.04±0.037

TABLE III
COMPARISON OF DUAL-DSA WITH BASELINES ON THYROID NODULE ULTRASOUND DATASETS. DICE (%).

Method	TNUI	TN3K	DDTI
Few-shot			
SENet	23.44±0.012	29.38±0.005	38.36±0.017
PANet	15.00±0.010	31.06±0.009	21.74±0.017
Interactive			
SAM-Med2D-B	3.69±0.006	18.79±0.005	11.41±0.016
SAM-Med2D-P	17.79±0.015	30.17±0.008	26.83±0.021
In-context			
UniverSeg	23.09±0.013	33.05±0.007	45.04±0.023
Tyche	35.70±0.017	**41.96**±0.006	42.55±0.019
SegGPT	37.07±0.011	17.28±0.014	41.48±0.026
Dual-DSA	**42.34**±0.012	28.77±0.014	**45.64**±0.025
Dual-DSA (rej.)	47.04±0.015	37.57±0.015	49.69±0.024

demonstrating its ability to effectively leverage limited support information. Performance further improves with additional supports ($k = 8, 16, 32$), although the gains become marginal as the support set increases.

Tables I-III report segmentation performance on the endoscopy, breast ultrasound, and thyroid ultrasound datasets. To ensure fair evaluation, we report full-coverage results without confidence-based rejection as the primary setting, while also presenting the post-rejection performance separately. Under full-coverage evaluation, Dual-DSA achieves the best Dice score on 8 out of 11 datasets. In particular, Dual-DSA consistently outperforms all baselines across the endoscopy datasets, achieving Dice scores of 65.98%, 59.29%, 52.29%, and 35.08% on Kvasir-SEG, CVC-Clinic, CVC-Colon, and ETIS, respectively (Table I). On breast and thyroid ultrasound datasets, Dual-DSA remains competitive and achieves the best performance on BUS-WHU, BUS-UCLM, TNUI, and DDTI, including Dice scores of 59.98% on BUS-WHU and 45.64% on DDTI (Tables II and III). Although Tyche obtains higher full-coverage Dice scores on BUS-BRA and TN3K, Dual-DSA consistently outperforms SegGPT across all datasets, indicating that the proposed dynamic support alignment and evidential fusion strengthen the underlying attention-based visual ICL formulation.

In addition to the full-coverage results, confidence-based rejection further improves the Dice scores of Dual-DSA by filtering low-confidence predictions. This highlights the additional benefit of Aggregate Evidential Confidence Scoring, which is analyzed in detail in Section V-C. Overall, these results show that Dual-DSA outperforms the best baseline on most datasets, consistently improves upon its attention-based visual ICL backbone, and achieves further gains through confidence-based rejection.

TABLE IV
ROBUSTNESS EVALUATION ON KVASIR-SEG UNDER (A) CONTROLLED SUPPORT VARIATIONS AND (B) RANDOM VERSUS SIMILARITY RETRIEVED SUPPORT SELECTION. DICE (%) WITH COEFFICIENT OF VARIATION (%) IN PARENTHESES. BOLD INDICATES THE BEST DICE SCORE AND THE LOWEST COEFFICIENT OF VARIATION.

(a) Controlled support variations

Method	Similar	Diff. Pos.	Diff. Shape	Diff. Size
UniverSeg	83.47 (6.35)	1.68 (68.92)	29.12 (79.36)	67.47 (22.52)
Tyche	90.10 **(2.92)**	7.73 (113.36)	39.79 (44.16)	83.72 (10.53)
SegGPT	82.95 (16.38)	88.46 (5.21)	61.10 (48.10)	81.17 (13.82)
Dual-DSA	**91.26** (4.01)	**94.33 (1.95)**	**83.61 (11.83)**	**91.09 (4.33)**

(b) Support selection: random vs. similarity retrieval

Method	Random	Sim-Retrieval
SegGPT	73.48 (30.16)	81.27 (22.74)
Dual-DSA	**91.15 (4.74)**	**93.16 (2.61)**

B. Robustness of Dual-DSA to Support Set Variations

Next, we evaluate the robustness of Dual-DSA under varying support sets. To assess adaptability, we construct support subsets with controlled variations in object position, shape, and size. As shown in Fig. 4, baseline ICL methods exhibit noticeable degradation in segmentation performance under such variations, with reasonable results observed only under similar support settings. In contrast, Dual-DSA maintains consistent and accurate segmentation performance across all support set settings. We further quantify robustness by repeatedly sampling support sets within each variation category and performing inference over 100 trials per method. Table IV(a) reports the mean Dice scores and corresponding coefficients of variation under these controlled support variations. Dual-DSA achieves the lowest variation across all categories, indicating more stable performance even when the support set is not optimally selected.

To compare Dual-DSA's intrinsic support integration with external retrieval-based support selection, we evaluate a simple similarity-based retrieval setting, denoted as Sim-Retrieval, where supports are selected using cosine similarity between image-level MedSAM encoder features. Table IV(b) shows that Sim-Retrieval improves both SegGPT and Dual-DSA over random support sampling, confirming the importance of support relevance in visual ICL. However, Dual-DSA remains substantially more accurate and stable than SegGPT under both support selection strategies. This suggests that external

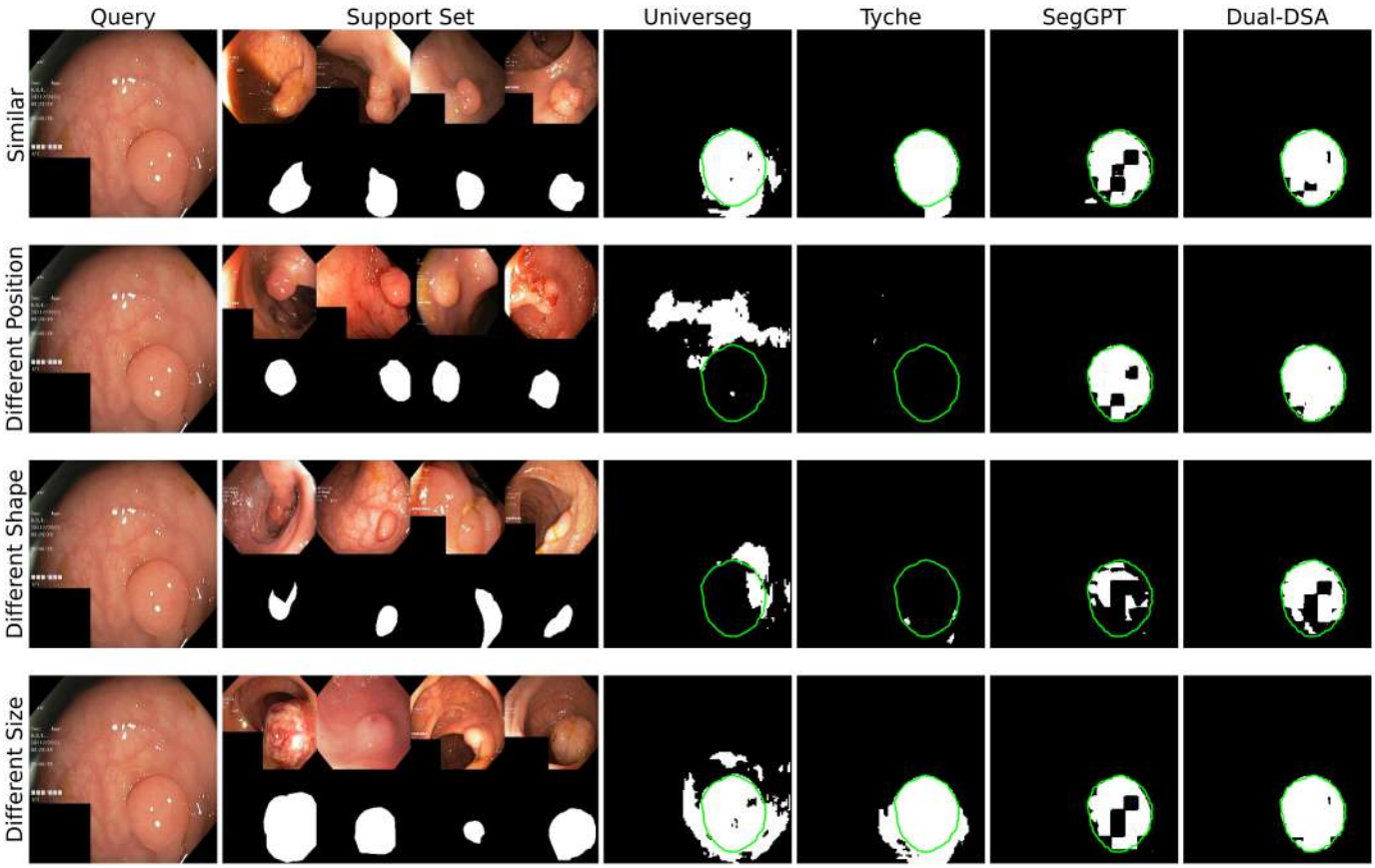

Fig. 4. Segmentation results of Dual-DSA and baselines (UniverSeg, Tyche, SegGPT) under different support set variations. Row 1: Similar Supports; row 2: Different Position; row 3: Different Shape; row 4: Different Size. Green contours denote ground truth segmentations.

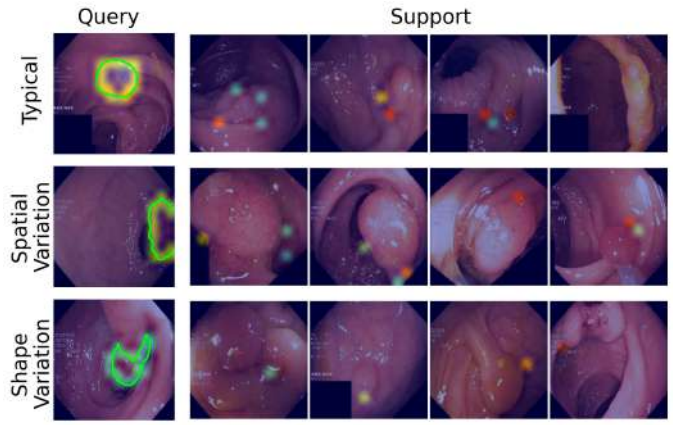

Fig. 5. Attention from query-boundary tokens to support regions for three cases: Typical, Spatial Variation, and Shape Variation. Boundary tokens contain both foreground and background pixels in the ground-truth masks, whose contours are shown on the queries.

retrieval can improve the support input space, but cannot intrinsically resolve support-set sensitivity. Dual-DSA instead addresses this limitation through intrinsic query-adaptive support integration.

To understand the source of Dual-DSA's robustness beyond support selection, we analyze the behavior of the proposed DynPatch-Align module. DynPatch-Align enables query tokens to attend to all support tokens via cross-attention, facilitating patch-level interactions across the entire support set. In Fig. 5, we visualize attention maps for query tokens corresponding to lesion boundaries, which are inherently ambiguous yet contain discriminative cues for segmentation. These tokens were identified from ground-truth mask patches containing both foreground and background pixels. The results show that boundary query patches selectively attend to spatially corresponding regions across multiple supports, particularly around object boundaries. The attention patterns vary across supports in both magnitude and spatial extent, indicating that each support contributes differently depending on its relevance to the query region. This behavior demonstrates that DynPatch-Align adaptively aggregates complementary information from multiple supports, rather than processing each support independently before global feature averaging, which cannot account for region-specific correspondences across supports. We further observe consistent query-to-support alignment even under large spatial and shape variations (Fig. 5, rows 2, 3), suggesting that patch-level interactions provide flexibility in handling appearance discrepancies.

Building on the empirical results in Table IV and the attention analysis in Fig. 5, we further contextualize the robustness of Dual-DSA with respect to related adaptive support strategies. Similarity-based retrieval or weighting strategies improve support relevance by modifying or reweighting the support input space, as reflected by the gains from Sim-

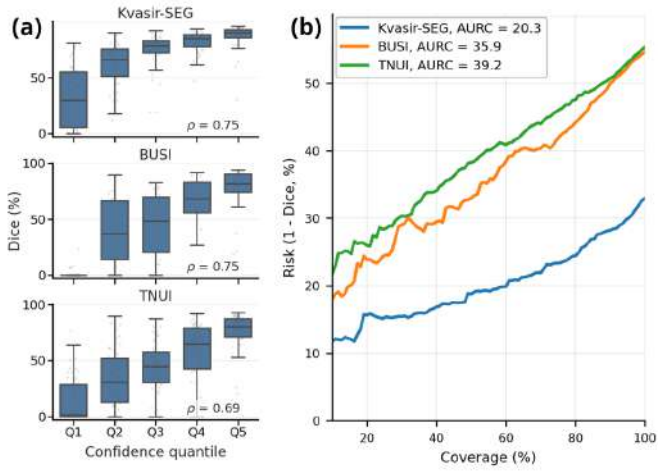

Fig. 6. Multi-dataset reliability analysis of Aggregate Evidential Confidence Scoring. (a) Dice distributions across confidence quantiles (Q1: lowest; Q5: highest) with Spearman ρ reporting the rank correlation between confidence and Dice. (b) Risk-coverage curves show selective prediction after sorting test samples by confidence, where risk is defined as 1-Dice and lower AURC indicates better reliability ranking.

TABLE V
RELIABILITY METRICS AND RETAINED DICE AT MATCHED COVERAGE.

Method	ECE-Dice (↓)	AURC (↓)	Dice@100% (↑)	Dice@80% (↑)
Endoscopy: Kvasir-SEG				
UniverSeg	50.12	69.27	31.03±0.224	31.74±0.235
Tyche	46.38	59.54	39.19±0.237	40.10±0.245
SegGPT	11.97	40.56	50.67±0.315	52.47±0.337
Dual-DSA	**10.10**	**20.28**	**67.05**±0.266	**75.98**±0.184
Breast US: BUSI				
UniverSeg	51.33	71.64	30.33±0.283	31.31±0.277
Tyche	45.74	56.34	43.60±0.267	44.33±0.274
SegGPT	34.50	82.47	28.06±0.333	26.63±0.347
Dual-DSA	**23.99**	**35.88**	**45.35**±0.343	**56.26**±0.296
Thyroid US: TNUI				
UniverSeg	51.10	72.67	24.73±0.265	26.73±0.275
Tyche	57.01	57.35	37.15±0.259	38.53±0.271
SegGPT	**28.70**	71.81	34.05±0.325	33.31±0.334
Dual-DSA	29.57	**39.19**	**44.67**±0.309	**51.77**±0.285

Retrieval in Table IV(b). However, such strategies primarily operate at the level of whole support images or global support weights, limiting their ability to handle spatially varying query–support correspondences or suppress irrelevant regions within an otherwise useful support. Stochastic support sampling improves robustness through exposure to diverse configurations during training, but performs support–query interactions independently for each support and relies on implicit averaging at inference time, lacking mechanisms to resolve uninformative or conflicting support evidence.

In contrast, Dual-DSA performs joint, query-conditioned patch-level alignment across all supports, coupled with uncertainty-aware aggregation, enabling spatially adaptive utilization of relevant regions and reduced influence of unreliable support evidence. Although performance may degrade in the absence of informative supports, the results in Fig. 4 show that Dual-DSA remains robust under practically challenging configurations.

C. Reliability Analysis via Aggregate Evidential Confidence Scoring

Here, we analyze the effectiveness of Aggregate Evidential Confidence Scoring in identifying poor quality samples. To provide a comprehensive reliability assessment, Fig. 6 evaluates the relationship between confidence and segmentation quality across representative datasets from three medical domains. As shown in Fig. 6(a), higher confidence quantiles correspond to higher Dice scores on Kvasir-SEG, BUSI, and TNUI, with strong positive Spearman rank correlations across all three datasets. This shows that the informativeness of the proposed confidence score is consistently observed across representative imaging domains.

Qualitatively, these poor quality cases correspond to clinical attributes that pose challenges for segmentation. From Fig. 7, cases in the lowest 20^{th} percentile of confidence scores frequently exhibit challenging attributes, including small lesions, low contrast, heterogeneous backgrounds, and acquisition artifacts such as motion blur. In contrast, cases belonging to the highest 20^{th} percentile contain larger, well-positioned polyps with clearer backgrounds and more distinct boundary delineation. This distinction is further reflected in the corresponding uncertainty maps. Low-confidence samples exhibit widespread or diffuse uncertainty, particularly in cases with small lesions, heterogeneous backgrounds, or motion artifacts, while low contrast cases show near-uniform uncertainty across the image. The absence of localized low-uncertainty regions leads to elevated global uncertainty, and consequently lower image-level confidence. In contrast, high-confidence cases show uncertainty primarily along lesion boundaries, yielding higher confidence after aggregation. These observations indicate that the proposed confidence scoring provides useful signals for identifying unreliable predictions, enabling downstream reliability-aware decision making.

To further evaluate the practical utility of the proposed confidence scoring, we adopt a selective prediction protocol under matched coverage. Specifically, predictions are ranked by confidence, and only the highest-confidence samples are retained for evaluation. Fig. 6(b) summarizes this behavior using risk-coverage curves, where risk is defined as $1 - \text{Dice}$. Across all three datasets, Dual-DSA achieves lower retained risk at reduced coverage levels, indicating that low-confidence predictions are associated with higher segmentation error and can be effectively prioritized for rejection.

Table V further reports calibration and selective prediction metrics on Kvasir-SEG, BUSI, and TNUI. For baselines without explicit uncertainty estimation, we derive confidence from the inverse of the mean pixel-wise entropy of the predicted probability map. ECE-Dice quantifies the calibration gap between image-level confidence and observed Dice, while AURC summarizes the risk-coverage curve.

Dual-DSA achieves lower or comparable ECE-Dice and the lowest AURC across the evaluated datasets, indicating stronger confidence-quality alignment and more effective confidence-

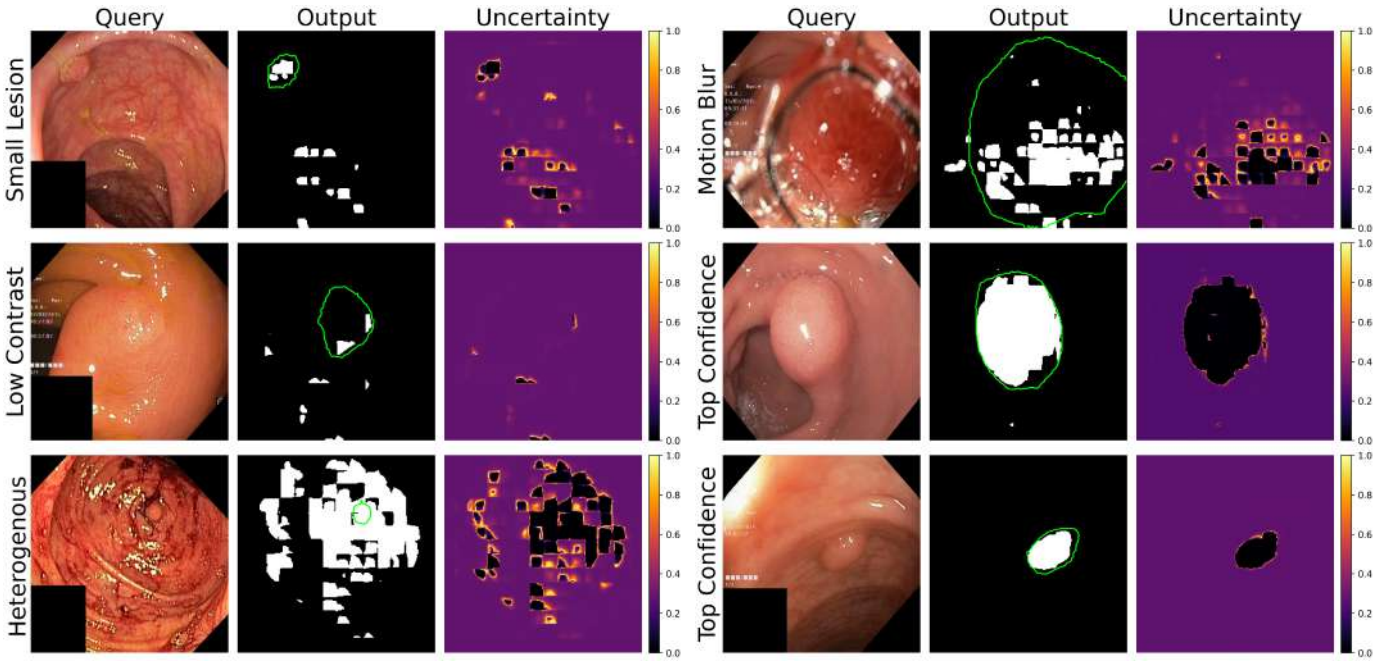

Fig. 7. Qualitative results showing predictions and uncertainty maps for challenging clinical cases. Left (rows 1-3): Small Lesion, Low Contrast, and Heterogeneous backgrounds; right (row 1): Motion Blur. For comparison, two high-confidence cases are shown (right; rows 2-3). Green contours denote ground truth segmentations.

TABLE VI
ABLATION OF DUAL-DSA COMPONENTS ACROSS THREE MEDICAL DOMAINS. DICE (%).

Configuration	Endoscopy	Breast US	Thyroid Nodule
Base	41.30±0.106	38.20±0.075	30.36±0.065
+ DynSuppPatch	44.19±0.112	46.08±0.086	36.52±0.069
+ DynSuppSet	51.08±0.128	46.80±0.077	34.41±0.104
Dual-DSA	**53.16**±0.121	**50.30**±0.067	**38.92**±0.075

based ranking of prediction reliability. This is further reflected in retained Dice at reduced coverage. When retaining the highest-confidence 80% of samples, Dual-DSA improves from 67.05% to 75.98% on Kvasir-SEG, 45.35% to 56.26% on BUSI, and 44.67% to 51.77% on TNUI (Table V). In contrast, the entropy-based confidence scores of other ICL baselines exhibit weaker or less consistent selective prediction performance. By explicitly modeling evidence and uncertainty, Dual-DSA provides confidence scores better aligned with segmentation reliability, supporting more effective confidence-aware inference and validating the reliability of its evidential confidence estimation.

D. Ablations

1) Contribution of Dual-DSA Components: We investigate the contribution of the two alignment components by implementing dynamic patch-level and set-level alignment separately. From Table VI, individually adding DynSuppPatch and DynSuppSet yields average Dice gains of 5.64 and 7.48 points, respectively, and combining both components improves over the base configuration by 10.84 points. These results demonstrate the complementary contributions of DynSuppPatch and DynSuppSet, supporting the proposed dual-level alignment framework.

2) Impact of Support Configuration: The performance and efficiency of Dual-DSA are influenced by the quantity of information in the support data. As shown in Fig. 8, we analyze Dice performance and inference time from two aspects: (1) the number of support images (left), and (2) the number of support sets (right). Increasing the number of support images improves segmentation accuracy across all three medical domains, with the most significant gains observed when increasing from $k = 2$ to 4 images. However, additional support images introduce non-trivial computational overhead, leading to a proportional increase in inference latency. A similar trend is observed for the number of support sets, where performance gains diminish beyond two sets, with Dice plateauing while latency continues to increase linearly. Based on these observations, we adopt four support images and two support sets as they provide the best trade-off between accuracy and computational efficiency.

3) Comparison of Support Set Fusion Strategies: Dual-DSA employs evidential fusion to integrate multiple support-conditioned query evidences for segmentation. In our implementation, we relax the independence assumption of Dempster-Shafer Theory (DST) by constructing two support sets from $k = 4$ images, allowing partial overlap between sets (Section IV-B). We compare Dual-DSA with alternative fusion strategies, including logit averaging and Dual-DSA (strict DST), which uses disjoint support sets to more closely follow the DST independence assumption (Table VII). Logit averaging generally underperforms compared to using all supports as a single set, indicating that simple aggregation is insufficient without explicit evidence modeling. In contrast, evidential fusion yields consistent improvements, demonstrating

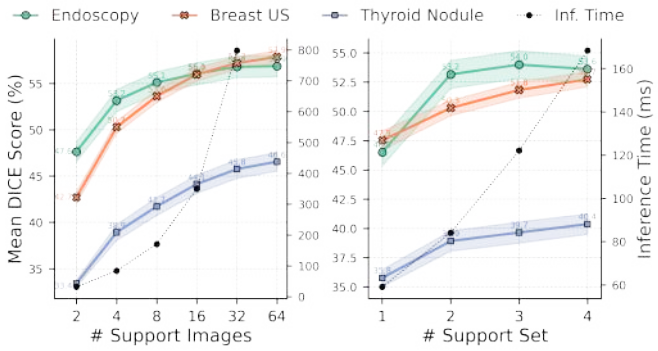

Fig. 8. Mean Dice (%) and inference time (ms) under varying support configurations. Left: number of support images. Right: number of support sets. Dice scores are aggregated across datasets within the same medical domain. Shaded regions denote standard errors. Inference time for 64 support images is omitted for clarity.

TABLE VII
COMPARISON OF SUPPORT SET FUSION STRATEGIES. DICE (%).

Configuration	Endoscopy	Breast US	Thyroid Nodule
No fusion	46.51±0.118	47.52±0.082	35.76±0.065
Logit averaging	49.70±0.121	44.18±0.081	33.66±0.103
Dual-DSA (strict DST)	**53.31**±0.114	48.40±0.068	37.02±0.054
Dual-DSA (relaxed DST)	53.16±0.121	**50.30**±0.067	**38.92**±0.075

that DynSuppSet provides reliability-aware integration beyond conventional prediction averaging. To assess the impact of the independence assumption, we partition the $k = 4$ supports into two disjoint sets of two images each. This removes support overlap, but also reduces the contextual information available to each support-conditioned prediction. While performance remains comparable, the overlapping formulation achieves modestly higher performance, supporting its use as a practical trade-off between support-set distinctness and contextual information.

E. Efficiency Analysis

Due to the additional cross-attention and evidential fusion operations, we analyze the computational overhead of Dual-DSA in terms of inference latency, FLOPs, and GPU memory usage. Table VIII summarizes the computational efficiency of Dual-DSA and baseline methods. Dual-DSA incurs higher computational cost than lightweight baselines, but achieves substantially improved segmentation accuracy. Despite increased latency, the method remains within practical runtime (84.08 ms per image) and is comparable to large-scale models such as SAM-Med2D, indicating a reasonable performance-efficiency trade-off for accuracy-critical applications.

VI. CONCLUSION

In this paper, we presented Dual-DSA, a novel visual ICL framework for medical image segmentation that addresses the fundamental limitation of support set sensitivity in ICL. Unlike existing ICL approaches that treat supports uniformly or rely on external retrieval mechanisms, Dual-DSA introduces intrinsic, query-adaptive support integration at

TABLE VIII
COMPUTATIONAL EFFICIENCY COMPARISON OF DUAL-DSA AND BASELINES. ± DENOTES STANDARD DEVIATION OVER 300 RUNS. MS, G, GB, AND M DENOTE MILLISECONDS, GIGA (FLOPS), GIGABYTES, AND MILLIONS, RESPECTIVELY.

Method	Inf. Time (ms)	FLOPs (G)	Max GPU (GB)	Param. (M)
SENet	4.35±10.23	12.0	0.16	0.88
PANet	40.97±8.56	32.1	0.06	14.7
SAM-Med2D-B	92.21±31.89	46.0	1.47	271
SAM-Med2D-P	95.77±36.86	46.0	1.47	271
UniverSeg	4.20±5.77	19.9	0.20	1.2
Tyche	5.78±4.62	33.4	0.19	1.7
SegGPT	30.39±1.00	689	1.81	370
Dual-DSA	84.08±0.73	1063	2.17	370

two complementary levels, combining spatially adaptive patch alignment with reliability-aware support-set fusion. Extensive experiments on 11 medical imaging datasets demonstrated that Dual-DSA achieves strong performance against state-of-the-art methods, obtaining the best Dice score on 8 out of 11 datasets. Furthermore, the derived confidence scores correlate well with segmentation accuracy, enabling selective rejection of unreliable predictions for quality control in clinical workflows. These capabilities make Dual-DSA particularly suitable for clinical deployment, where annotation scarcity and the need for trustworthy predictions are critical concerns. By enabling robust adaptation to new target domains without additional fine-tuning and providing built-in quality assurance, our framework offers a practical pathway toward reliable AI-assisted medical image analysis.

This work has several limitations that require further investigation. First, while cross-attention and evidential fusion introduce additional computational overhead, our empirical analysis shows that the increase remains manageable under moderate support sizes, though it may still pose challenges in highly resource-constrained clinical environments. Second, we evaluated our method only on binary segmentation tasks using two-dimensional medical images. The evidential formulation could be extended to multi-class segmentation by modeling class-wise evidence across $C > 2$ channels, together with class-aware confidence aggregation. Separately, extending Dual-DSA to three-dimensional volumetric segmentation would require adapting its support-conditioning and alignment mechanisms while managing memory and computational costs. Future work will explore lightweight architectures, multi-class and three-dimensional segmentation, and active learning strategies to reduce labeling burden.

REFERENCES

[1] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, 2022.

[2] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of U-Net," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[3] Z. Wang, Y. Wang, Y. Feng, J. Du, Y. Liu, R. S. M. Goh, and L. Zhen, "Continuous disentangled joint space learning for domain generalization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 7, pp. 13 122–13 134, 2025.

[4] J. S. Yoon, K. Oh, Y. Shin, M. A. Mazurowski, and H.-I. Suk, "Domain generalization for medical image analysis: A review," *Proc. IEEE*, 2024.

[5] H. Guan and M. Liu, "Domain adaptation for medical image analysis: A survey," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, 2022.

[6] S. Hu, Z. Liao, J. Zhang, and Y. Xia, "Domain and content adaptive convolution based multi-source domain generalization for medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 42, no. 1, pp. 233–244, 2022.

[7] Y. Feng, Z. Wang, X. Xu, Y. Wang, H. Fu, S. Li, L. Zhen, X. Lei, Y. Cui, J. S. Z. Ting *et al.*, "Contrastive domain adaptation with consistency match for automated pneumonia diagnosis," *Medical Image Analysis*, vol. 83, p. 102664, 2023.

[8] Y. Feng, X. Xu, H. Fu, Y. Wang, Z. Wang, L. Zhen, R. S. M. Goh, and Y. Liu, "Category-aware test-time training domain adaptation," in *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2024, pp. 300–306.

[9] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "UniverSeg: Universal medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[10] A. Bar, Y. Gandelsman, T. Darrell, A. Globerson, and A. A. Efros, "Visual prompting via image inpainting," in *Advances in Neural Information Processing Systems*, 2022.

[11] X. Wang, W. Wang, Y. Cao, C. Shen, and T. Huang, "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*, 2023, pp. 6830–6839.

[12] Y. Zhang, K. Zhou, and Z. Liu, "What makes good examples for visual in-context learning?" in *Advances in Neural Information Processing Systems*, 2023, pp. 17 773–17 794.

[13] J. Gao, Q. Lao, Q. Kang, P. Liu, C. Du, K. Li, and L. Zhang, "Boosting your context by dual similarity checkup for in-context learning medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 44, no. 1, pp. 310–319, 2025.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3992–4003.

[15] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nat. Commun.*, vol. 15, no. 1, p. 654, 2024.

[16] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang, H. Sun, J. He, S. Zhang, M. Zhu, and Y. Qiao, "SAM-Med2D," *CoRR*, vol. abs/2308.16184, 2020. [Online]. Available: https://arxiv.org/abs/2308.16184

[17] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang, "SegGPT: Towards segmenting everything in context," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1130–1140.

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

[19] O. Rubin, J. Herzig, and J. Berant, "Learning to retrieve prompts for in-context learning," in *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 2655–2671.

[20] W. Zhou, Y. E. Jiang, R. Cotterell, and M. Sachan, "Efficient prompting via dynamic in-context learning," *CoRR*, vol. abs/2305.11170, 2023. [Online]. Available: https://arxiv.org/abs/2305.11170

[21] S. Cai, T. Mensah-Boateng, X. Kuksov, J. Yuan, and S. Tang, "The power of adaptation: Boosting in-context learning through adaptive prompting," *CoRR*, vol. abs/2412.17891, 2024. [Online]. Available: https://arxiv.org/abs/2412.17891

[22] Z. Yang, D. Dai, P. Wang, and Z. Sui, "Not all demonstration examples are equally beneficial: Reweighting demonstration examples for in-context learning," in *Findings of the Association for Computational Linguistics: EMNLP*, 2023, pp. 13 209–13 221.

[23] J. U. Allingham, J. Ren, M. W. Dusenberry, X. Gu, Y. Cui, D. Tran, J. Z. Liu, and B. Lakshminarayanan, "A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models," in *Proceedings of the International Conference on Machine Learning*, 2023, pp. 26–47.

[24] C. Huang, L. Huang, and J. Huang, "Divide, reweight, and conquer: A logit arithmetic approach for in-context learning," *CoRR*, vol. abs/2410.10074, 2024. [Online]. Available: https://arxiv.org/abs/2410.10074

[25] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in Neural Information Processing Systems*, 2018, p. 3183–3193.

[26] A. Malinin and M. Gales, "Uncertainty estimation in autoregressive structured prediction," in *Proceedings of the International Conference on Learning Representations*, 2021.

[27] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[28] W. Bao, Q. Yu, and Y. Kong, "Evidential deep learning for open set action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[29] K. Holmquist, L. Klasén, and M. Felsberg, "Evidential deep learning for class-incremental semantic segmentation," in *Image Analysis*, 2023, pp. 32–48.

[30] K. Zou, X. Yuan, X. Shen, M. Wang, and H. Fu, "Tbrats: Trusted brain tumor segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 503–513.

[31] M. Feng, K. Xu, N. Wu, W. Huang, Y. Bai, Y. Wang, C. Wang, and H. Wang, "Trusted multi-scale classification framework for whole slide image," *Biomed. Signal Process. Control*, vol. 89, p. 105790, 2024.

[32] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," *CoRR*, vol. abs/1911.07069, 2019. [Online]. Available: https://arxiv.org/abs/1911.07069

[33] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imaging Graph.*, vol. 43, pp. 99–111, 2015.

[34] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Trans. Med. Imaging*, vol. 35, no. 2, pp. 630–644, 2016.

[35] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, 2014.

[36] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief*, vol. 28, p. 104863, 2020.

[37] W. Gómez-Flores, M. J. Gregorio-Calas, and W. C. D. A. Pereira, "Bus-bra: A breast ultrasound dataset for assessing computer-aided diagnosis systems," *Med. Phys.*, vol. 51, no. 4, pp. 3110–3123, 2024.

[38] N. Vallez, G. Bueno, O. Deniz, M. A. Rienda, and C. Pastor, "Bus-uclm: Breast ultrasound lesion segmentation dataset," 2024.

[39] J. Huang, J. Zhang, Y. Zhang, X. Li, X. Ma, J. Deng, H. Shen, D. Wang, L. Mei, and C. Lei, "Busi_whu: Breast cancer ultrasound image dataset," 2023.

[40] X. Nie, X. Zhou, T. Tong, X. Lin, L. Wang, H. Zheng, J. Li, E. Xue, S. Chen, M. Zheng, C. Chen, and M. Du, "N-net: A novel dense fully convolutional neural network for thyroid nodule segmentation," *Front. Neurosci.*, vol. 16, p. 872601, 2022.

[41] H. Gong, G. Chen, R. Wang, X. Xie, M. Mao, Y. Yu, F. Chen, and G. Li, "Multi-task learning for thyroid nodule segmentation with thyroid region prior," in *Proc. IEEE Int. Symp. Biomed. Imaging*, 2021, pp. 257–261.

[42] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero, "An open access thyroid ultrasound image database," in *Proc. Int. Symp. Med. Inf. Process. Anal.*, 2015, p. 92870W.

[43] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "Squeeze & Excite guided few-shot segmentation of volumetric images," *Med. Image Anal.*, vol. 59, p. 101587, Oct. 2019.

[44] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2019.

[45] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proceedings of the International Conference on Learning Representations*, 2022.

[46] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecol.*, vol. 26, no. 3, pp. 297–302, 1945.