Medical Image Analysis 104 (2025) 103639

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

An orchestration learning framework for ultrasound imaging: Prompt-Guided Hyper-Perception and Attention-Matching Downstream Synchronization

Zehui Lin [a], Shuo Li [b], Shanshan Wang [c], Zhifan Gao [d], Yue Sun [a], Chan-Tong Lam [a], Xindi Hu [e], Xin Yang [f], Dong Ni [f], Tao Tan [a,*]

[a] Faculty of Applied Sciences, Macao Polytechnic University, Macao Special Administrative Region of China
[b] Department of Biomedical Engineering and the Department of Computer and Data Science, Case Western Reserve University, Cleveland, OH, USA
[c] Paul C. Lauterbur Research Center for Biomedical Imaging, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China
[d] School of Biomedical Engineering, Sun Yat-sen University, Guangzhou, 510275, China
[e] Shenzhen RayShape Medical Technology Co. Ltd., Shenzhen, 518071, China
[f] School of Biomedical Engineering, Shenzhen University, Shenzhen, 518060, China

ARTICLE INFO

Keywords:
Ultrasound imaging
Orchestration learning
Downstream Synchronization
Prompt Guided learning

ABSTRACT

Ultrasound imaging is pivotal in clinical diagnostics due to its affordability, portability, safety, real-time capability, and non-invasive nature. It is widely utilized for examining various organs, such as the breast, thyroid, ovary, cardiac, and more. However, the manual interpretation and annotation of ultrasound images are time-consuming and prone to variability among physicians. While single-task artificial intelligence (AI) solutions have been explored, they are not ideal for scaling AI applications in medical imaging. Foundation models, although a trending solution, often struggle with real-world medical datasets due to factors such as noise, variability, and the incapability of flexibly aligning prior knowledge with task adaptation. To address these limitations, we propose an orchestration learning framework named PerceptGuide for general-purpose ultrasound classification and segmentation. Our framework incorporates a novel orchestration mechanism based on prompted hyper-perception, which adapts to the diverse inductive biases required by different ultrasound datasets. Unlike self-supervised pre-trained models, which require extensive fine-tuning, our approach leverages supervised pre-training to directly capture task-relevant features, providing a stronger foundation for multi-task and multi-organ ultrasound imaging. To support this research, we compiled a large-scale Multi-task, Multi-organ public ultrasound dataset (M^2-US), featuring images from 9 organs and 16 datasets, encompassing both classification and segmentation tasks. Our approach employs four specific prompts—Object, Task, Input, and Position—to guide the model, ensuring task-specific adaptability. Additionally, a downstream synchronization training stage is introduced to fine-tune the model for new data, significantly improving generalization capabilities and enabling real-world applications. Experimental results demonstrate the robustness and versatility of our framework in handling multi-task and multi-organ ultrasound image processing, outperforming both specialist models and existing general AI solutions. Compared to specialist models, our method improves segmentation from 82.26% to 86.45%, classification from 71.30% to 79.08%, while also significantly reducing model parameters.

1. Introduction

Ultrasound imaging has become an essential tool in clinical diagnostics due to its affordability, portability, safety (being radiation-free), real-time capability, and non-invasive nature. It is particularly prevalent in the examination of superficial organs such as the breast (Spak et al., 2017) and thyroid (Tessler et al., 2017). Beyond these, ultrasound is extensively used in a wide range of applications including ovarian evaluations (Wu et al., 2018), cardiac function assessments (Folland et al., 1979), prenatal fetal monitoring (He et al., 2021), carotid artery examinations (Stein et al., 2008), liver (Ferraioli and Monteiro, 2019) and kidney function tests (Mostbeck et al., 2001), and appendicitis diagnosis (Mostbeck et al., 2016), among others. Despite its widespread

* Corresponding author.
E-mail address: taotan@mpu.edu.mo (T. Tan).

https://doi.org/10.1016/j.media.2025.103639
Received 22 January 2025; Received in revised form 9 April 2025; Accepted 9 May 2025
Available online 27 May 2025

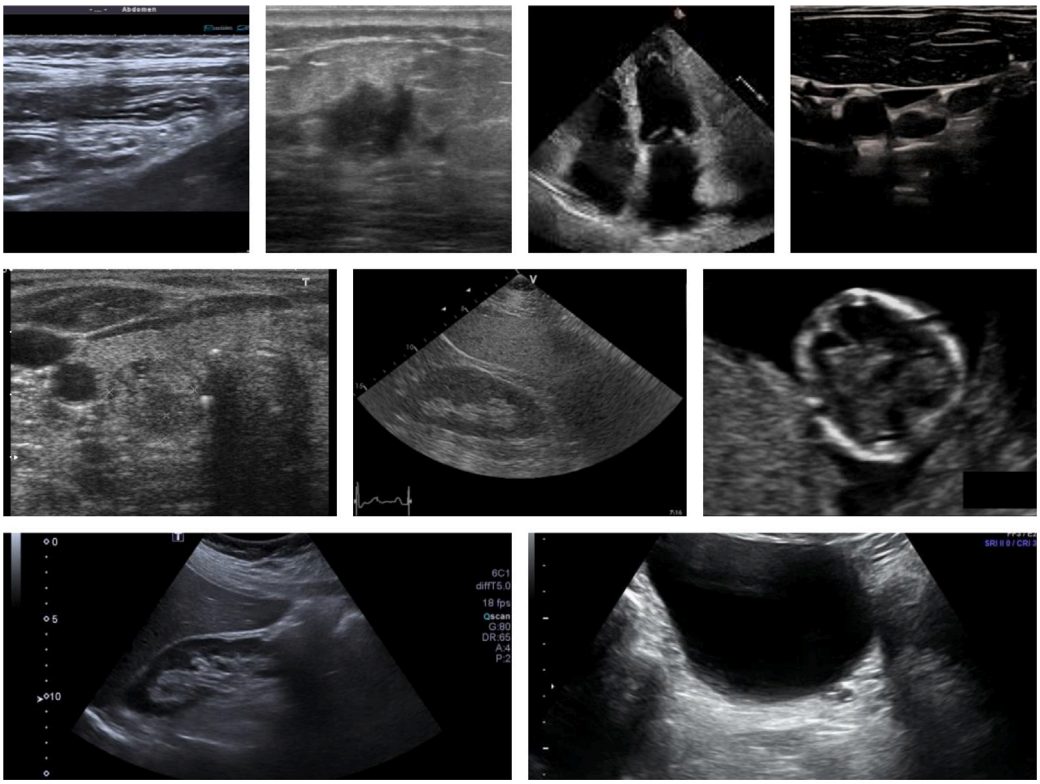

Fig. 1. Examples of ultrasound images demonstrating the complexity and diversity of ultrasound imaging as well as the differences between various organs. From left to right, top to bottom: appendix image, breast image, cardiac image, carotid artery image, thyroid image, liver image, fetal head image, kidney image and ovarian image.

usage, traditional ultrasound image analysis models are often developed for specific tasks or organ-specific datasets. These single-purpose models lack flexibility and generalization, making it challenging to scale across the diverse range of ultrasound imaging scenarios encountered in clinical practice. Furthermore, deploying multiple single-purpose models increases computational and storage demands, limiting their practicality in real-world applications. This highlights the need for a unified, automated solution capable of handling various tasks and organs efficiently, thereby enhancing diagnostic accuracy and consistency.

Existing self-supervised pretraining ultrasound foundation models are often less efficient compared to supervised pretraining, requiring a vast amount of data and significant computational resources. Furthermore, these models typically require extensive task-specific fine-tuning during deployment, whereas supervised pretraining can often be applied directly. Although supervised pretraining is a relatively efficient approach, it still faces several challenges. Firstly, as shown in Fig. 1, traditional ultrasound image analysis models are usually trained on specific and limited datasets. These models are designed to handle a particular type of ultrasound analysis problem, which restricts their scope and makes them incapable of effectively dealing with the diverse ultrasound images in clinical practice. Moreover, when deploying such models on devices, the use of single-purpose models requires the deployment of multiple models, thereby imposing limitations on storage and memory. Secondly, in the development of an orchestration multi-organ ultrasound model, one has to confront the complexity and diversity of ultrasound images. There are significant variations in tumor images and anatomical structures, and the image scales can also be inconsistent. The differences among ultrasound images from different body parts make unified processing arduous. Thirdly, many existing general-purpose medical image models encounter difficulties during fine-tuning on new data. Fine-tuning often results in catastrophic forgetting, causing the model to lose previously acquired knowledge and leading to poor generalization abilities. These challenges clearly indicate the necessity for innovative solutions to construct robust and versatile ultrasound image analysis models.

1.1. Related work

The challenges in developing an orchestration ultrasound image model are not unique and are also present in many other modalities and datasets. Numerous studies have proposed solutions to these challenges.

1.1.1. Self-supervised learning

One significant approach is self-supervised learning, which involves designing an appropriate proxy task to learn patterns from large-scale unlabeled data and then transferring this knowledge to downstream tasks. For example, Wu et al. (2023) proposed BROW, a foundational model for whole-slide images based on self-distillation, which improves performance through multi-scale input and color augmentation, excelling in various downstream tasks. Li et al. (2023) developed D-LMBmap, an automatic deep learning pipeline for whole-brain neural circuit analysis in mice, utilizing video transformers and self-supervised learning to achieve efficient and accurate brain analysis. Wang et al. (2023b) introduced Endo-FM, a foundational model for endoscopy video analysis, achieving good performance in classification, segmentation, and detection tasks through large-scale self-supervised pre-training and a unique spatio-temporal matching strategy. Zhou et al. (2023) presented RETFound, a foundational model for retinal images based on self-supervised learning, which adapts effectively to various disease detection tasks, excelling in both ocular and systemic disease detection and prediction. Hua et al. (2023) proposed Patho-Duet, a foundational model for pathology slide analysis covering H&E and IHC images, validating its efficiency and generalization capability in multiple downstream tasks through a new self-supervised learning framework and two proxy tasks. Wang et al. (2023d) introduced a

Table 1
Current progress of general model for medical image analysis.

Reference	Model/Benchmark/Dataset name	Methods
Wu et al. (2023)	BROW	Self-Supervised Pretraining
Li et al. (2023)	D-LMBmap	
Wang et al. (2023b)	Endo-FM	
Zhou et al. (2023)	RETFound	
Hua et al. (2023)	PathoDuet	
Wang et al. (2023d)	PCT-Net	
Jiao et al. (2024)	USFM	
Kang et al. (2023)	Deblurring MAE	
Huang et al. (2023a)	A-Eval	Data Benchmark
Wang et al. (2023c)	MedFMC	
Ding et al. (2023)	SNOW	Data Generation
Qin et al. (2022)	VLMs for Medical	Visual-Text Modeling
Zhang et al. (2023)	CITE	
Wang et al. (2023a)	SAM-Med3D	Large-Scale Supervised Pretraining
Cheng et al. (2023)	SAM-Med2D	
Lei et al. (2023)	MedLSAM	
Huang et al. (2023b)	STU-Net	
Lin et al. (2023)	SAMUS	
Ma et al. (2024)	MedSAM	

self-supervised learning method based on Volume Fusion and Parallel Convolution and Transformer Network (PCT-Net) for pre-training 3D medical image segmentation, showing excellent performance in multiple downstream segmentation tasks. Jiao et al. (2024) proposed USFM, a universal ultrasound foundational model, by building a large-scale multi-organ, multi-center, multi-device ultrasound database and using a spatial-frequency dual-masking image modeling method for self-supervised pre-training, demonstrating good generality, performance, and label efficiency in various downstream tasks. Kang et al. (2023) introduced Deblurring MAE, a method that incorporates a deblurring task into masked autoencoder (MAE) (He et al., 2022) pre-training, enhancing the ability of MAE to recover details in ultrasound images, thereby improving its performance in ultrasound image recognition tasks. While self-supervised learning (SSL) offers the significant advantage of leveraging vast unlabeled datasets, thereby avoiding costly manual annotation, the features learned via proxy tasks often require substantial fine-tuning to effectively adapt to specific downstream medical tasks, particularly when domain shifts exist (Jiao et al., 2024). Conversely, supervised pre-training (SL) on diverse, labeled datasets like M^2-US, despite its annotation cost, directly optimizes representations for the target tasks (*e.g.*, segmentation, classification). For developing a unified framework like PerceptGuide, intended for robust performance across multiple known tasks and organs with potentially minimal downstream adaptation, this direct task alignment learned via SL can provide a more efficient pathway to deployment effectiveness. Furthermore, SSL and supervised approaches like ours are not mutually exclusive; future work could explore leveraging upstream SSL pre-training for rich feature extraction followed by supervised training with PerceptGuide to instill task-specific knowledge and prompt-based orchestration.

1.1.2. Data diversity

According to Zhang and Metaxas (2023), data diversity is one of the key factors in training foundational models. Therefore, some researchers have conducted in-depth studies on dataset construction or generation. Huang et al. (2023a) introduced the A-Eval benchmark for evaluating the cross-dataset generalization capability of abdominal multi-organ segmentation models, emphasizing the importance of data diversity and model size on generalization. Wang et al. (2023c) proposed a new dataset and benchmark, MedFMC, for evaluating foundational models in medical image classification, covering a variety of real-world clinical tasks and different medical image modalities, validating the effectiveness and limitations of some foundational models in medical image classification tasks through experiments. Ding et al. (2023) introduced a large-scale synthetic pathology image dataset, SNOW, for breast cancer segmentation, demonstrating its effectiveness and competitiveness in model training through quality validation, enhancing nuclear segmentation performance. From these works, we understand that training an effective foundational model requires sufficiently diverse data.

1.1.3. Multimodal learning

Regarding data, some researchers have proposed utilizing text data to construct multimodal visual-text models combining text (reports) and images. Qin et al. (2022) explored how to use large-scale pre-trained visual language models (such as GLIP Li et al., 2022) for medical image understanding, improving model performance in medical image detection and classification tasks by manually designing effective medical prompts and automatically generating medical prompts, showing that pre-trained Vision Language Models (VLMs) can be effectively transferred to the medical domain through prompt learning. Zhang et al. (2023) proposed the CITE framework, which enhances pathology image classification performance by injecting textual knowledge into the foundational model adaptation through linking image and text embeddings, demonstrating strong model extensibility and excellent performance under data-limited conditions. Training both the text encoder and image encoder simultaneously in such models is challenging due to the significant differences in their feature spaces, and these models face the issue of small datasets, as paired medical image-text datasets are currently accumulating and are much less abundant than natural image-text paired datasets. However, incorporating structured semantic information can alleviate this issue to some extent, as it does not rely heavily on large volumes of paired data, thus mitigating the data scarcity problem.

1.1.4. Supervised learning

Unlike the approach of self-supervised learning pre-training followed by transfer to downstream tasks (Table 1), another significant approach for general models is to first construct a rich dataset and then conduct large-scale supervised pre-training, allowing the model to be used directly. Wang et al. (2023a) proposed SAM-Med3D, a model that modifies Segment Anything Model (SAM) (Kirillov et al., 2023) with 3D positional encoding for 3D medical image segmentation, displaying excellent performance and generalization on multiple datasets. Cheng et al. (2023) fine-tuned SAM on a large-scale medical image dataset to obtain SAM-Med2D, significantly improving various medical image segmentation tasks and showing excellent performance and generalization in segmenting different anatomical structures, modalities, and organs. Lei et al. (2023) proposed MedLSAM, a fully automated

SAM medical adaptation model, including MedLAM for 3D medical image localization and SAM for segmentation, reducing annotation workload and exhibiting good performance. Huang et al. (2023b) designed a series of scalable and transferable medical image segmentation models, STU-Net, demonstrating strong performance and transferability in different downstream tasks through supervised pre-training on a large-scale dataset. Lin et al. (2023) introduced SAMUS and its end-to-end version AutoSAMUS, general models based on SAM for ultrasound image segmentation, achieving better segmentation performance and automatic segmentation capability through improvements such as CNN branches, feature adapters, and positional adapters, as well as an automatic prompt generator (APG). Ma et al. (2024) proposed MedSAM, a foundational model for medical image segmentation, trained and fine-tuned on a large-scale medical image dataset, exhibiting superior performance and generalization in various segmentation tasks compared to existing models. While models like SAM (Kirillov et al., 2023) and its medical adaptations (*e.g.*, SAM-Med2D Cheng et al., 2023, MedSAM Ma et al., 2024) have demonstrated potential through spatial prompts (points, boxes), these approaches primarily focus on localization cues and may lack the semantic richness required to address ultrasound-specific challenges. Ultrasound imaging is inherently complex due to factors such as significant noise, artifacts (*e.g.*, acoustic shadowing, enhancement), and high inter-patient variability in anatomical structures across acquisition conditions.

This approach is practical and can effectively avoid the shortcomings of self-supervised learning design. However, currently, there is no general supervised large-scale pre-training method for the ultrasound modality that can simultaneously solve multiple tasks of classification and segmentation. In multi-task learning, it is common to encounter inconsistencies between task objectives (*i.e.*, conflicts between tasks). Google's HydraNet (Mullapudi et al., 2018) addresses this issue by sharing underlying features while employing task-specific heads, effectively mitigating such conflicts. This can be seen as an implementation of "co-learning", where a shared training mechanism enables multiple tasks to optimize their respective objectives simultaneously. However, in contrast to our approach, HydraNet has a notable limitation. It requires separate segmentation and classification decoders (seg/cls decoders) for different organs. For instance, it would need one set of seg/cls decoders for breast images and another for thyroid images. In our proposed framework, we take advantage of the flexibility of prompts. Regardless of the type of organ's ultrasound image, we can utilize the same seg decoder or the same cls decoder through prompt adaptation. This unique design of network structure reuse in our method allows for greater efficiency and simplicity compared to HydraNet. This paper follows our own approach to conduct research on ultrasound data.

1.2. Contributions

This work represents a significant extension of our previous conference paper UniUSNet (Lin et al., 2024). Notably, this paper extends UniUSNet by introducing a hyper-perception module with prompts and an attention-matching downstream synchronization stage, while expanding the dataset from 7 to 9 organs (9.7k images to 33k images) and adding comprehensive evaluations (including SOTA, thorough ablation and failure case analysis). To address the aforementioned challenges, this paper proposes an orchestration learning framework **PerceptGuide** for ultrasound classification and segmentation, which incorporates a hyper-perception module and leverages the characteristics of ultrasound prompts to enhance flexibility and tunability. Our contributions can be summarized as follows:

- **Comprehensive Dataset Collection**: We have compiled a large-scale public ultrasound dataset named M^2-US (**M**ulti-position **M**ulti-task). This dataset includes ultrasound images of 9 different organs or body parts, covering a total of 16 datasets (6 classification and 10 segmentation datasets), with over 33,000 images.
- **Prompt-Guided Hyper-Perception Orchestration Ultrasound Framework**: We propose a general framework for ultrasound image processing that addresses multi-task and multi-organ ultrasound image classification and segmentation. The core of this framework is the hyper-perception module we developed, which introduces model prompts. This not only enhances the model's flexibility but also incorporates prior knowledge. We designed four specific prompts based on the intrinsic properties of ultrasound images: Object, Task, Input, and Position.
- **Attention-Matching Downstream Synchronization Stage**: For the extension purpose, we introduce a downstream synchronization training stage that effectively fine-tunes our hyper-perception module as an adapter for new data, thereby improving the model's generalization capability.

The rest of the paper is organized as follows: In Section 2, we detail our methodology, including the categories of model prompts, the hyper-perception module, and the downstream synchronization stage. Section 3 covers our experimental protocol, where the M^2-US dataset is described in detail, including data preprocessing, implementation details, and evaluation metrics. In Section 4, we present our experimental results and corresponding analysis and discussion. Finally, Section 5 concludes the whole paper.

2. Methodology

As illustrated in Fig. 2, our proposed PerceptGuide is an orchestration learning framework for ultrasound imaging, handling both segmentation and classification. It starts by mapping prompts (Object, Task, Input, and Position) into embeddings for decoders. An input image is encoded by ViT and then processed by hyper-decoders. For new data, a downstream synchronization stage tunes the network by fine-tuning prompt embedding layers while freezing others.

2.1. Network architecture in orchestration learning

In the PerceptGuide framework, we adopt Swin-Unet as the backbone network. The Vision Transformer (ViT) encoder within it plays a crucial role in transforming the input ultrasound image into image embeddings. These image embeddings then serve as the input for the subsequent decoders. There are two main decoders in our framework: the hyper-segmentation decoder and the hyper-classification decoder. The hyper-segmentation decoder is responsible for generating segmentation outputs, while the hyper-classification decoder is designed to produce predicted classification labels. Each decoder contains multiple transformer blocks, and in each block, the hyper-perception module is integrated to handle the prompts effectively. During the forward pass, the image embeddings flow through the decoders, and the final outputs are obtained based on the task requirements. In addition to the encoder and decoders, there are fully connected layers involved in generating the prompt embeddings. These fully connected layers map the prompt vectors to the corresponding embeddings, which are then used in the hyper-perception module to adjust the attention weights. Overall, this architecture enables the framework to perform ultrasound image classification and segmentation tasks efficiently and adapt to different datasets through the downstream synchronization stage.

2.2. Prompt categories: A holistic approach to upstream and downstream tasks

In recent advancements of general segmentation models for natural images, such as the SAM, it has been demonstrated that incorporating appropriate prompt knowledge can significantly enhance network accuracy. Here, we posit that the improvement in network performance due to prompts can be understood from two perspectives. Firstly, prompts are highly flexible, meaning that the same network with the same input

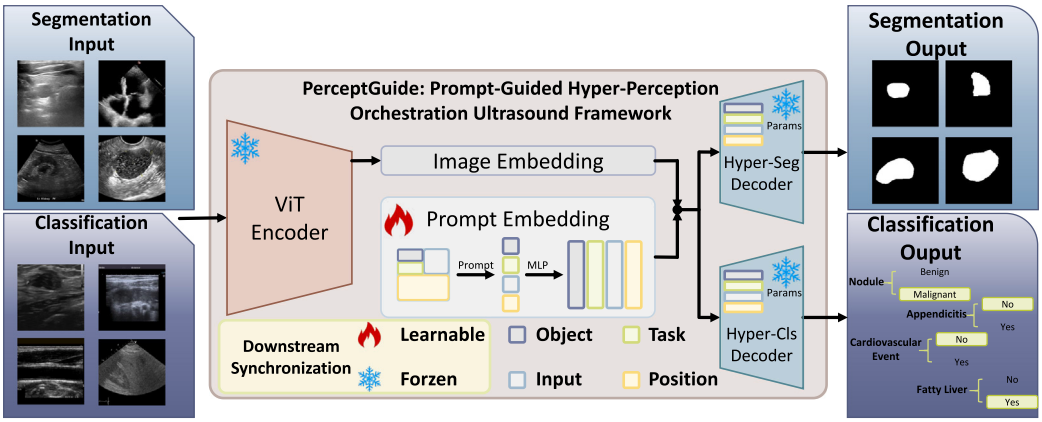

Fig. 2. This figure illustrates the framework of our proposed network, named PerceptGuide. We utilize Swin-Unet (Cao et al., 2022) as the backbone. In the classification decoder, there are skip connections similar to those in the segmentation decoder, but no upsampling layers. The hyper-perception operations are present in each block of the transformer in every layer of the decoders. During the downstream synchronization stage, we fine-tune only the MLP layers involved in generating prompt embeddings.

image can produce different outputs based on the given prompts. For instance, in an image of a thyroid with a tumor, prompting for the gland will yield a segmentation map of the gland, while prompting for the tumor will result in a segmentation map of the tumor. Secondly, prompts can be viewed as a form of prior knowledge, which provides the network with certain priors. For example, organ-related prompts can supply crucial prior information by explicitly indicating the organ type (*e.g.*, thyroid, liver, or kidney), guiding the model to focus on relevant anatomical structures and features specific to that organ.

As shown in Fig. 4(a), we propose four types of prompts: Object, Task, Input, and Position. Unlike SAM, our network handles two tasks simultaneously, so our prompts are not merely simple points or boxes as in SAM (manual intervention required) but are instead category-specific prompts with richer semantic information and advantages of full automation. The introduction of such prompts not only brings prior knowledge that enhances the network's inductive bias but also improves the network's flexibility. Specifically, the Object and Input prompts enable the network to handle both tumors and organs with different types of inputs. Meanwhile, the Task and Position prompts help the network learn different organs more effectively by leveraging prior knowledge. The following sections provide a detailed description of each type of prompt.

2.2.1. Object Prompt

Object Prompt includes two optional prompts: *Tumor* and *Organ*. Due to the different characteristics of tumors and organs, tumors are often irregular and focus on heterogeneity, while organs are regular and focus on continuity and structure. Therefore, we need different prompts to handle these two different inputs, making our network capable of flexibly adapting to the specific features of tumors and organs.

2.2.2. Task Prompt

Task Prompt includes two optional prompts: *Segmentation* and *Classification*. We know that the information required for classification and segmentation is different. Classification often needs to focus more on overall features, while segmentation requires more attention to details related to localization. Although we already have dedicated decoders for segmentation and classification, the decoders are not inherently designed to identify which specific task they should prioritize. By guiding the decoders with task prompts, we provide an inductive bias in addition to the ground truth supervision signal, further guiding the network to learn the task. This gradient and supervision signal can also be backpropagated to the encoder, allowing it to perform image encoding tasks accordingly, thus enabling our network to leverage prior knowledge to different tasks.

2.2.3. Input Prompt

Input Prompt includes three types: *Whole*, *Local*, and *Highlighted*. As illustrated in Fig. 3, these prompts are designed to handle different input scenarios, such as whole images, locally magnified regions, and situations where prior location information (via masks) is available for lesions or anatomical structures. These prompts help the network learn more effectively from diverse inputs. By using these prompts, we can not only improve the network's ability to recognize images of different spatial granularities but also maximize the use of available information.

As shown in both 4 (a) and the detailed process in Fig. 3, *Whole* represents the input of the entire original image. *Local* represents the input of a cropped local part of the image, based on the minimum bounding rectangle of the available mask. (In actual implementation, the cropped image is resized and padded to match the size of the *Whole* input for consistency.) *Highlighted* represents the whole image with highlighted prompts overlaid onto the lesion or anatomical area defined by the mask, allowing the network to better focus on these regions. This is particularly beneficial for ultrasound images, which often suffer from significant noise; highlighting the informative region, as suggested by Choe and Shim (2019), can help suppress the influence of background noise on the model's classification judgment.

It is important to note the source and usage of masks for generating *Local* and *Highlighted* prompts. While in our current training methodology we utilize the ground truth segmentation masks provided within the datasets, it is conceptually feasible to generate these masks using initial segmentation predictions from the model itself or another segmentation tool, offering flexibility in application scenarios.

Note that during training, for the segmentation task, we use *Whole* and *Local* prompts. We do not use *Highlighted* prompts for the hyper-seg decoder, as illustrated in Fig. 3, to avoid label information leakage. In the classification task, all three prompts (*Whole*, *Local*, and *Highlighted*) are used, enabling our network to flexibly handle different spatial granularities and leverage prior location information when available.

2.2.4. Position Prompt

Position Prompt includes ten types: *Ovary*, *Breast*, *Thyroid*, *Carotid*, *Appendix*, *Liver*, *Cardiac*, *Head*, *Kidney* and *Others*. Here, prompts for different parts introduce different domain knowledge to the network. Introducing different prompts is equivalent to switching the network's inductive bias to utilize knowledge specific to a particular location to handle different inputs. Additionally, we introduce an "*Others*" prompt to handle inputs that do not belong to the previous nine parts. With these prompts, our network can leverage prior knowledge to adapt to a wide range of anatomical locations.

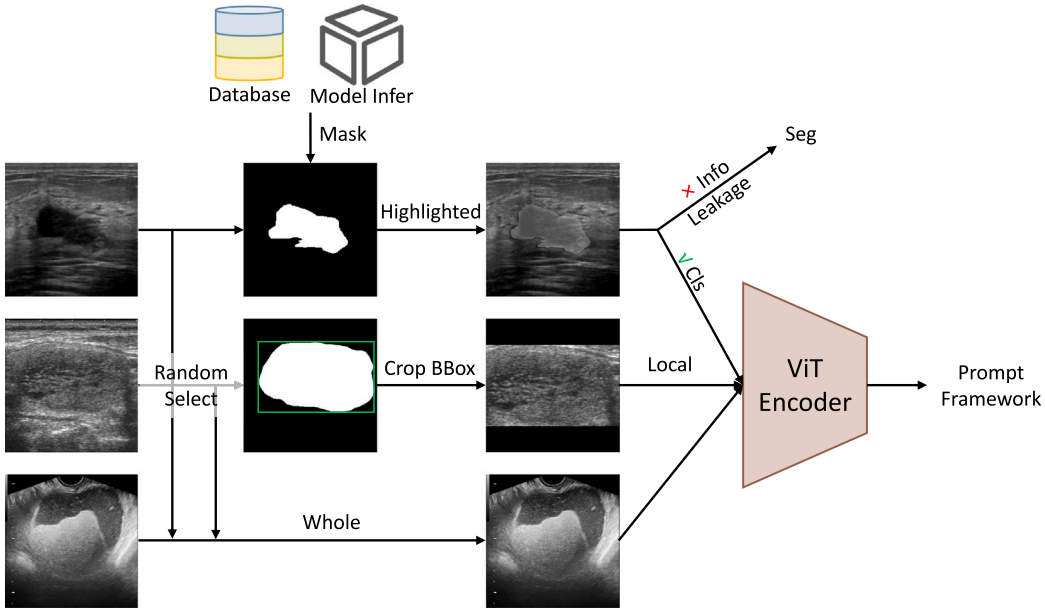

Fig. 3. Illustration of the generation process for the three input prompt types: *Whole*, *Local*, and *Highlighted*. Starting from an original ultrasound image, the *Whole* prompt uses the image directly. If a segmentation mask (sourced from ground truth or potentially predicted) is available, the *Local* prompt is generated by cropping the image based on the mask's bounding box, while the *Highlighted* prompt overlays the mask information onto the original whole image. Note the task-specific usage: the *Highlighted* prompt is used only for classification to avoid information leakage during segmentation training.

2.2.5. Ultrasound-specific design rationale

The four prompt categories are deliberately designed to address key challenges in ultrasound imaging:

- Noise & Artifacts: The Input prompt (*Whole/Local/Highlighted*) allows the model to adaptively focus on regions of interest while suppressing artifacts. For example, a "*Highlighted*" input directs attention to lesion areas even amidst acoustic shadows.
- Anatomical Variability: The Position prompt injects organ-specific priors (*e.g.*, ovarian follicular patterns vs. thyroid nodular textures), mitigating the impact of inter-organ structural differences.
- Task Ambiguity: The Task prompt disentangles feature learning objectives—classification requires holistic context awareness, whereas segmentation demands pixel-level precision.
- Pathological Diversity: The Object prompt (*Tumor/Organ*) differentiates between irregular malignant masses and regular anatomical structures, addressing heterogeneity in lesion appearances.

2.3. Prompt strategy – Hyper-Perception module

The attention mechanism is the most common and crucial part of the transformer architecture, enabling feature interaction and integration. The core formula of attention is as follows:

$$Z = \text{softmax}(\frac{QK^T}{\sqrt{d}})V, \tag{1}$$

where Q, K, and V represent the query, key, and value, and d is the feature dimension, respectively. The significance of this formula is that for each query, we calculate its similarity with all keys and then use this similarity to perform a weighted sum of the values. However, this attention mechanism is relatively fixed and cannot adjust its weights according to different inputs. In many general-purpose large models, especially in the Segmentation Anything Model (SAM), the prompts used are usually generated directly by the prompt encoder. This method of joining is relatively indirect and lacks direct interaction with the network.

We propose a novel hyper-perception module designed to integrate prompts effectively. The core of this module is to transform prompt embeddings and hyper-perception parameters into weight matrices, which are then incorporated as computational weights within the attention mechanism. This hyper-perception module replaces the standard attention module in the decoder, embedding within each transformer layer and each transformer block. This approach addresses the inflexibility of traditional attention mechanisms and allows explicit interaction with the network, thereby better guiding the network's learning process.

As illustrated in Fig. 4(b), the four types of prompts mentioned earlier are represented by the vectors i_α, i_β, i_γ, and i_δ (i_α represents the object prompt, i_β the task prompt, i_γ the input prompt, and i_δ the position prompt). These vectors are one-hot vectors indicating different prompt options ($i_\cdot \in \mathbb{R}^{k_\cdot \times 1}$, where $k_\cdot$ is the number of options for that type of prompt). These vectors are first mapped through an MLP. We use four independent MLPs with non-shared weights, denoted as f_α, f_β, f_γ, and f_δ, each processing a different type of prompt. Through these MLP layers, we obtain the prompt embeddings for each type of prompt. The specific formulas are as follows:

$$x_\alpha = f_\alpha(i_\alpha), \tag{2}$$

$$x_\beta = f_\beta(i_\beta), \tag{3}$$

$$x_\gamma = f_\gamma(i_\gamma), \tag{4}$$

$$x_\delta = f_\delta(i_\delta), \tag{5}$$

where $x_\cdot \in \mathbb{R}^{d \times 1}$, and d is the feature dimension of the transformer block, which varies between different blocks.

After constructing the prompt embeddings, each transformer block also includes a set of hyper-perception learnable parameters with dimensions d, denoted as c_α, c_β, c_γ, and c_δ ($c_\cdot \in \mathbb{R}^{1 \times d}$). Before computing the attention in each transformer block, we perform a matrix multiplication between the prompt embeddings and the hyper-perception parameters. The resulting formulas are:

$$\omega_\alpha = x_\alpha \times c_\alpha, \tag{6}$$

$$\omega_\beta = x_\beta \times c_\beta, \tag{7}$$

$$\omega_\gamma = x_\gamma \times c_\gamma, \tag{8}$$

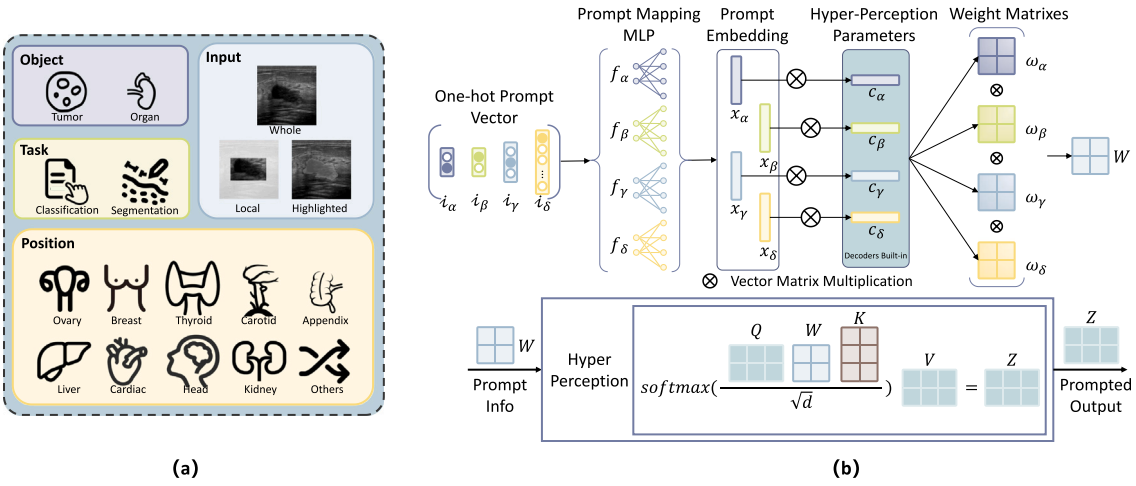

Fig. 4. Overview of the Prompt Content and Hyper-Perception Module. (a) Four primary categories of prompts, encompassing a total of 17 different prompt. (b) Workflow of the hyper-perception module (labels correspond to components described in Section 2.3). Prompt embeddings x, derived via MLPs f from Prompt vectors i, are multiplied with corresponding learnable parameters c to yield prompt-specific weight matrices ω. These matrices are subsequently fused via matrix multiplication (Eq. (10)) into a unified weight matrix W. This matrix W then directly modulates the standard attention calculation (Eq. (11)) within the decoder layers, integrating the specified multi-faceted prompt guidance into the attention mechanism.

$$\omega_\delta = x_\delta \times c_\delta. \tag{9}$$

After matrix multiplication, we obtain four weight matrices related to the prompts, denoted as ω_α, ω_β, ω_γ, and ω_δ ($\omega_\cdot \in \mathbb{R}^{d\times d}$). We then fuse these weight matrices into a unified weight matrix W for the four types of prompts through matrix multiplication, as follows:

$$W = \omega_\alpha \times \omega_\beta \times \omega_\gamma \times \omega_\delta \quad (W \in \mathbb{R}^{d\times d}) \tag{10}$$

Finally, we embed this unified weight matrix W into the attention mechanism, performing a parameterized weighted attention mechanism with the following formula:

$$Z = \text{softmax}(\frac{QWK^T}{\sqrt{d}})V, \tag{11}$$

where $Z \in \mathbb{R}^{n\times d}$, n is the sequence length and d is the feature dimension. In this way, we obtain a hyper-perception module that incorporates prompts. This hyper-perception module is embedded into each transformer block, better guiding the network's learning. The detailed procedure to compute the prompt embeddings, weight matrices, and the unified weight matrix W is outlined in Algorithm 1. Moreover, this module is very lightweight, requiring only a few additional parameters and computations. The learnable internal hyper-perception parameters enable the network to incorporate the knowledge and preferences indicated by the prompts. Additionally, the prompt embedding generating process is generated by MLPs, which facilitates subsequent fine-tuning.

Algorithm 1 Hyper-Perception Module with Prompt Strategy (see Fig. 4 for details)

Input: Query Q, Key K, Value V, Prompt vectors $i_\alpha, i_\beta, i_\gamma, i_\delta$, Learnable parameters $c_\alpha, c_\beta, c_\gamma, c_\delta$.
Output: Attention output Z.
Step 1: Generate Prompt Embeddings.
For each $i_\cdot \in \{i_\alpha, i_\beta, i_\gamma, i_\delta\}$: Map $i_\cdot$ to prompt embedding $x_\cdot$ using corresponding MLP $f_\cdot$. $x_\cdot = f_\cdot(i_\cdot)$.
Step 2: Compute Weight Matrices.
For each $x_\cdot \in \{x_\alpha, x_\beta, x_\gamma, x_\delta\}$: Compute weight matrix $\omega_\cdot$ by matrix multiplication with learnable parameter $c_\cdot$. $\omega_\cdot = x_\cdot \cdot c_\cdot$.
Step 3: Fuse Weight Matrices.
Combine the four weight matrices $\omega_\alpha, \omega_\beta, \omega_\gamma, \omega_\delta$ into a unified weight matrix W. $W = \omega_\alpha \cdot \omega_\beta \cdot \omega_\gamma \cdot \omega_\delta$.
Step 4: Apply Parameterized Weighted Attention.
Integrate W into the attention mechanism. $Z = \text{softmax}\left(\frac{QWK^T}{\sqrt{d}}\right)V$.
Output: The attention output Z.

2.4. Attention-Matching Downstream Synchronization Stage

A model trained on a large dataset should support efficient fine-tuning rather than being trained directly on a new dataset. Training a model from scratch on a new dataset can lead to catastrophic forgetting and incur high training costs. Fine-tuning allows a model to leverage the vast knowledge acquired during pre-training while effectively adapting to new datasets. By incorporating small, fine-tunable modules into the network, the model can learn dataset-specific features without significantly altering the primary pre-trained weights. This approach to adapting can significantly reduce the number of parameters required for fine-tuning, preserving the generalization capabilities of the original pre-trained model. This results in less storage space and faster training times, making the process more efficient and scalable.

Our downstream synchronization method is designed to achieve efficient fine-tuning, as illustrated in Fig. 2. During the downstream synchronization stage, we freeze the parameters of the ViT encoder and the two hyper-decoders (hyper-seg and hyper-cls). Instead of retraining the entire model, we only fine-tune the MLP layers responsible for generating prompt embeddings on the new data. This allows our network to adapt to new datasets at a low cost. Importantly, the hyper-perception parameters, which are used in the hyper-decoders to compute hyper-perception, are also frozen and not fine-tuned. This decision is based on the belief that these parameters embody the prior knowledge learned by the network, which should remain unchanged. On the other hand, fine-tuning the fully connected (fc) layers is crucial because they directly modify the prompt embeddings, influencing the hyper-perception calculation weights. This adjustment allows for a degree of domain adaptation, ensuring that the network can effectively handle new data while retaining the benefits of its pre-trained knowledge.

2.5. Loss function and data sampling strategy

Our network employs two different loss functions for segmentation and classification tasks. For the segmentation task, we use a mixed loss

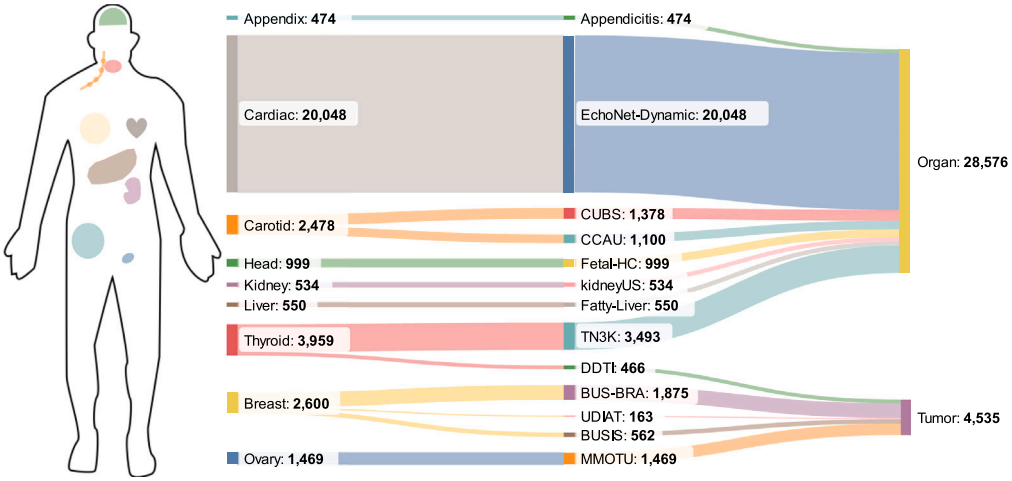

Fig. 5. Overview of the M^2-US dataset, showing the distribution of datasets across different organs, with a breakdown into organ and tumor images along with their respective quantities.

function composed of Cross-Entropy (CE) and Dice Loss, weighted at λ_1 and λ_2 respectively. For the classification task, we use the Cross-Entropy loss function. The loss functions can be expressed as follows:

$$L_{Seg} = \lambda_1 \times L_{CE} + \lambda_2 \times L_{Dice}, \tag{12}$$

$$L_{Cls} = L_{CE}, \tag{13}$$

$$L_{Total} = L_{Seg} + L_{Cls}. \tag{14}$$

In addition, we employ a sampling strategy to maintain balanced data distribution among different organs. This strategy is widely used in multi-organ training work (Jiao et al., 2024; Zhao et al., 2023). Since different organs represent different domain knowledge, it is crucial for the model to balance this knowledge effectively. If the number of samples for each organ is denoted as N_{organ}, the sampling probability weight for each organ's data is given by:

$$w_{organ} = \frac{1}{\sqrt{N_{organ}}}. \tag{15}$$

This sampling method helps preserve the original data distribution to some extent, unlike direct upsampling or downsampling to equal quantities, which can flatten the data distribution. Furthermore, we utilize extensive image augmentation strategies, including random rotation, random scaling, and horizontal flipping, to increase data diversity and enhance the model's generalization capability.

3. Experimental setup

3.1. M^2-US dataset and external validation datasets

A large-scale and comprehensive ultrasound image dataset is fundamental for developing a highly generalizable orchestration ultrasound model. In our research, we propose the M^2-US dataset, which is currently the largest publicly available **M**ulti-task, **M**ulti-organ US dataset. As shown in Fig. 5, the M^2-US dataset includes images from 9 organs, comprising 16 public datasets with a total of 33,111 ultrasound images. Among these, 4,535 images are tumor images, while the remaining 28,576 images depict organs or anatomical structures. As detailed in Table 2, the labeling varies across the 16 datasets; some datasets are classification tasks, some are segmentation tasks, and some include both. Specifically, there are 6 classification tasks and 10 segmentation tasks. Unlike some current related works, the dataset used in our experiments is entirely composed of publicly available data, without utilizing any private images. This ensures that our experimental results are reproducible and provides a benchmark for future research in this field, which is one of our contributions. Moreover, since our dataset is compiled from multiple distinct public datasets, it inherently includes multi-center and multi-machine scenarios, thereby enhancing the generalizability of our model. To ensure reproducibility, detailed information regarding the construction of the M^2-US dataset from its public sources, along with our experimental code and pre-trained models, are made available at our public repository.[1] By merging these datasets to construct our M^2-US, we can better train our model to achieve superior generalization capabilities.

In addition to our proposed M^2-US dataset used for internal validation, we selected several datasets for external validation, specifically to test the effectiveness of our method in the downstream synchronization stage. As shown in Table 3, we utilized three datasets for external validation: BUSI, HMC-QU, and TG3K. These datasets span three organs, two types of objects, and comprise a total of 6714 images, with both segmentation and classification labels.

3.2. Data preprocessing

The image dimensions vary across different datasets. Therefore, we first resized all images by proportionally adjusting the shorter side to 224 pixels, followed by a center crop to obtain 224 × 224 images. This approach prevents distortion of image content while ensuring consistent image size. All datasets were divided into training, validation, and test sets, with the training set comprising 70%, the validation set 10%, and the test set 20%. Some datasets, such as EchoNet-Dynamic, MMOTU, TG3K, and TN3K, come with predefined splits which we adhered to. Images from the same patient were placed in the same training or test set to prevent data leakage. Three data augmentation techniques were applied to the training set: random horizontal flip (probability of 0.5), random scaling (0.8x to 1.2x), and random rotation (angles from −20° to 20°).

We also standardized the dataset labels. Classification was unified as binary, and segmentation labels were harmonized as foreground and background. Most datasets conformed to this, such as the appendicitis classification in the Appendicitis dataset or the benign vs. malignant classification in the breast dataset. Two datasets required specific explanation: in the kidneyUS dataset, the segmentation label used was the contour of the entire kidney, and in the CUBS dataset, we predicted the occurrence of risk events.

[1] https://github.com/Zehui-Lin/PerceptGuide

Table 2
Summary of the M^2-US dataset characteristics. This table outlines 16 datasets, specifying task type (Seg/Cls/Both), organ, and image count. Class ratios (*e.g.*, (a:b)) are given for classification tasks. Totals for annotations per task, overall images, and overall annotations are provided in the final row.

Dataset	Segmentation	Classification	Organ	Image num
Appendicitis (Marcinkevičs et al., 2023)		✓ (174:300)	Appendix	474
BUS-BRA (Gómez-Flores et al., 2024)	✓	✓ (1268:607)	Breast	2600
BUSIS (Zhang et al., 2022)	✓			
UDIAT (Yap et al., 2017)	✓	✓ (109:54)		
CCAU (Momot, 2022)	✓		Carotid	2478
CUBS (Meiburger et al., 2022)		✓ (1124:254)		
DDTI (Pedraza et al., 2015)	✓		Thyroid	3959
TN3K (Gong et al., 2023)	✓	✓ (2283:1210)		
EchoNet-Dynamic (Ouyang et al., 2020)	✓		Cardiac	20 048
Fatty-Liver (Byra et al., 2018)		✓ (170:380)	Liver	550
Fetal-HC (van den Heuvel et al., 2018)	✓		Head	999
MMOTU (Zhao et al., 2022)	✓		Ovary	1469
kidneyUS (Singla et al., 2023)	✓		Kidney	534
Annotation Num	30 709	7933		Total Image: 33 111 Total Annotation: 38 642

Table 3
Summary of external validation datasets used in the study.

Dataset	Position	Object	Image number	Label
BUSI (Al-Dhabyani et al., 2020)	Breast	Tumor	780	Classification/Segmentation
HMC-QU (Degerli et al., 2021)	Cardiac	Organ	2349	Segmentation
TG3K (Wunderling et al., 2017)	Thyroid	Organ	3585	Segmentation

3.3. Implementation details

In this section, we provide more details about the implementation. We used the PyTorch framework (Paszke et al., 2019) to implement our model. The initial learning rate was set to 1×10^{-4}, followed by an exponential decay learning rate schedule. The Adam optimizer (Kingma and Ba, 2014) was employed for training. A batch size of 32 was used. The hyperparameter λ_1 and λ_2 in Eq. (12) were set to 0.4 and 0.6, respectively. This weighting for the combined Cross-Entropy and Dice loss was determined empirically on a validation subset of M^2-US and aligns with common practices in medical image segmentation aimed at balancing distribution matching and boundary accuracy, similar to approaches used in Isensee et al. (2021), Cao et al. (2022). Our model was trained on a server with 4 NVIDIA A4000 GPUs (16G) for 200 epochs, taking approximately 25 h. In addition, as shown in Fig. 5, the distribution of our data across organs is not balanced (with the cardiac organ comprising about 60% and the kidney organ only about 1.4%). Therefore, throughout the training process, we utilized the sampling strategy mentioned in Section 2.5 to mitigate the data imbalance issue.

3.4. Evaluation metrics

In this section, we detail the evaluation metrics used in our experiments. For classification tasks, accuracy was the primary metric. Accuracy is defined as the ratio of correctly predicted instances to the total instances and is calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

For segmentation tasks, we used the Dice coefficient, also known as the Dice similarity index, which is a measure of overlap between the predicted and ground truth segmentation. The Dice coefficient is given by:

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

The Dice coefficient ranges from 0 to 1, where 1 indicates perfect agreement between the predicted and ground truth segments.

4. Results and discussion

In Section 4.1, we present a comprehensive comparison of our method with other state-of-the-art (SOTA) approaches. Section 4.2 details ablation studies, where we evaluate the impact of removing the prompt guidance (*i.e.* Hyper-Perception module) based on the orchestration learning framework, and compare against different organ specialist models. Following an introduction to internal validation on the M^2-US dataset, we extend our analysis to external validation in Section 4.3. Section 4.4 offers insightful visualizations that provide deeper understanding of the results. Finally, we discuss the potential clinical applications of our model, address possible limitations, and suggest future research directions.

4.1. Comparison with state-of-the-art methods

Due to the absence of large-scale supervised pre-trained models for simultaneous multi-task classification and segmentation, we compared our approach with several popular general 2D image segmentation models. As shown in Table 4, these models include SAM, SAMUS, SAM-Med2D, and MedSAM. Our comparison focuses on these relevant state-of-the-art generalist foundation models adapted for medical imaging and strong specialist baselines. We compare against variants of the Segment Anything Model (SAM) (Kirillov et al., 2023), including MedSAM (Ma et al., 2024) and SAM-Med2D (Cheng et al., 2023).

Table 4
Comparison of our method with State-of-the-Art Orchestration models on M^2-US internal validation. Dice scores are averages calculated across the test sets of all segmentation datasets within the M^2-US dataset.

Model	Mode	Prompt	Prompt type	Dice
SAM	zero-shot	Box		76.44%
SAM	zero-shot	Point		37.88%
SAMUS	fine-tune	Point		86.36%
SAM-Med2D	zero-shot	Box	Interactive	71.76%
SAM-Med2D	zero-shot	Point	(Need Manual Delineation)	23.98%
SAM-Med2D	fine-tune	Box		73.91%
SAM-Med2D	fine-tune	Point		72.39%
MedSAM	zero-shot	Box		81.54%
MedSAM	fine-tune	Box		62.86%
Ours	–	Proposed 4 prompts	Automatic (Semantic prompt)	**92.75%**

Notably, we also consider SAMUS (Lin et al., 2023), a SAM adaptation specifically developed and evaluated for ultrasound segmentation. The SAMUS authors demonstrated its competitiveness against other advanced segmentation networks (*e.g.*, H2Former He et al., 2023, FAT-Net Wu et al., 2022, TransFuse Zhang et al., 2021) on ultrasound data (Lin et al., 2023), positioning it as a relevant SOTA benchmark for generalist ultrasound segmentation. Since SAMUS, SAM-Med2D, and MedSAM are SAM's adaptations for medical images, and SAMUS even specifically for ultrasound, we only used SAM in zero-shot mode as it is more commonly employed. SAMUS only supports point prompts and does not release its model weights, so we trained it from scratch on our M^2-US dataset using their open-source code. SAM-Med2D supports both Box and Point prompts; we tested these using their published weights in zero-shot mode and further fine-tuned them on our dataset. MedSAM, which only supports Box prompts, was tested in zero-shot mode and fine-tuned using their released checkpoint. For fairness, all SAM methods and our method were tested on cropped datasets using the minimum enclosing bounding boxes of the provided ground truth masks, in order to reduce any bias caused by pixel-level prompts. A key advantage of our approach is that our prompts are fully automated, eliminating the need for manual input like points or boxes from physicians or other models, setting it apart from semi-automated, interactive methods.

Analyzing the results in Table 4, SAM (Box, Zero-shot) achieved a Dice score of 76.44%, which is moderate compared to other methods, highlighting the domain gap due to pre-training on natural images. In contrast, SAM (Point, Zero-shot) scored 37.88%, significantly lower than others, possibly due to noise interference in ultrasound images affecting point prompts. SAMUS (Point, Fine-tune) achieved 86.36%, demonstrating improved performance, likely due to optimization for ultrasound images. SAM-Med2D (Box, Zero-shot) scored 71.76%, slightly lower than SAM (Box, Zero-shot), suggesting some adaptation but still limited by the domain gap. However, SAM-Med2D (Point, Zero-shot) at 23.98% is the lowest score, indicating severe issues with point prompts in noisy ultrasound data. SAM-Med2D (Box, Fine-tune) improved to 73.91%, showing that fine-tuning helps, but still not optimal. SAM-Med2D (Point, Fine-tune) scored 72.39%, much better than its zero-shot counterpart, but still not competitive. MedSAM (Box, Zero-shot) achieved 81.54% Dice score, attesting to its strong generalization capability derived from medical image pre-training. Notably, MedSAM (Box, Fine-tune) experienced a significant performance decline to 62.86%. Several factors could contribute to such drops during fine-tuning, including domain shifts between datasets, hyperparameter sensitivity, or catastrophic forgetting (CF). Given the large size of the MedSAM model and the substantial performance decrease observed when fine-tuning all parameters, we hypothesize that catastrophic forgetting – the overwriting of general pre-trained features critical for robustness when adapting to new data distributions – is the most likely primary factor here. This underscores a common challenge in adapting large foundation models, which our downstream synchronization strategy aims to mitigate by selectively fine-tuning only specific layers. Our method achieved the highest score of 92.75%, showcasing the effectiveness of our proposed four prompt modes, significantly outperforming all other methods. Our approach leads in performance by leveraging high-level semantic prompts, providing fully automated segmentation and reducing the need for expert intervention. Our method achieved the highest score of 92.75%, showcasing the effectiveness of our proposed four prompt modes, significantly outperforming all other methods. Our approach leads in performance by leveraging high-level semantic prompts, providing fully automated segmentation and reducing the need for expert intervention.

Furthermore, as detailed in our ablation studies (Section 4.2, Table 5), we also benchmark against 'Specialist' models trained individually for each task and dataset. These specialist models utilize a Swin-Unet backbone (Cao et al., 2022), a powerful transformer-based architecture widely adopted for medical image analysis and shown to be effective in various ultrasound applications, including breast (Yang and Yang, 2023) and thyroid (Sun et al., 2024) imaging, thus representing strong task-specific performance levels. The superior average performance of PerceptGuide over these strong specialist baselines further validates its effectiveness in the context of a unified multi-task, multi-organ framework.

We use radar charts to compare the performance of our model against SOTA models across different datasets. This serves as a supplement to Table 4, which only presents average results. As shown in Fig. 6, our model exhibits superior performance on most datasets compared to the SOTA models. The table already indicates that our model outperforms the SOTA models in average segmentation performance (92.75%). The radar charts further validate this by demonstrating that our high average performance is achieved through consistently strong results across all datasets. Some models show imbalanced performance, excelling in certain datasets while underperforming in others (*e.g.*, SAM-Med2D-Zero-shot-Box on UDIAT and Fetal-HC). This highlights our model's comprehensive capabilities and strong generalizability.

4.2. Ablation studies

We presents a comparison of ablation study results. As shown in Table 5, We compare the orchestration learning framework performance without prompt guidance (*i.e.*, without hyper-perception) and with various specialist models (single dataset, single model) to highlight the impact. Without prompt guidance, the framework remains the same, but the prompts are omitted. All datasets are mixed together for training using the multi-branch framework based on Swin-Unet. The table lists the parameter count for each model and the evaluation metrics. As previously mentioned, segmentation uses Dice, and classification uses accuracy. The best results for each dataset are bolded, and the best average results are underlined.

Analyzing the results, our model outperforms others in most datasets, especially in average results, leading in both segmentation

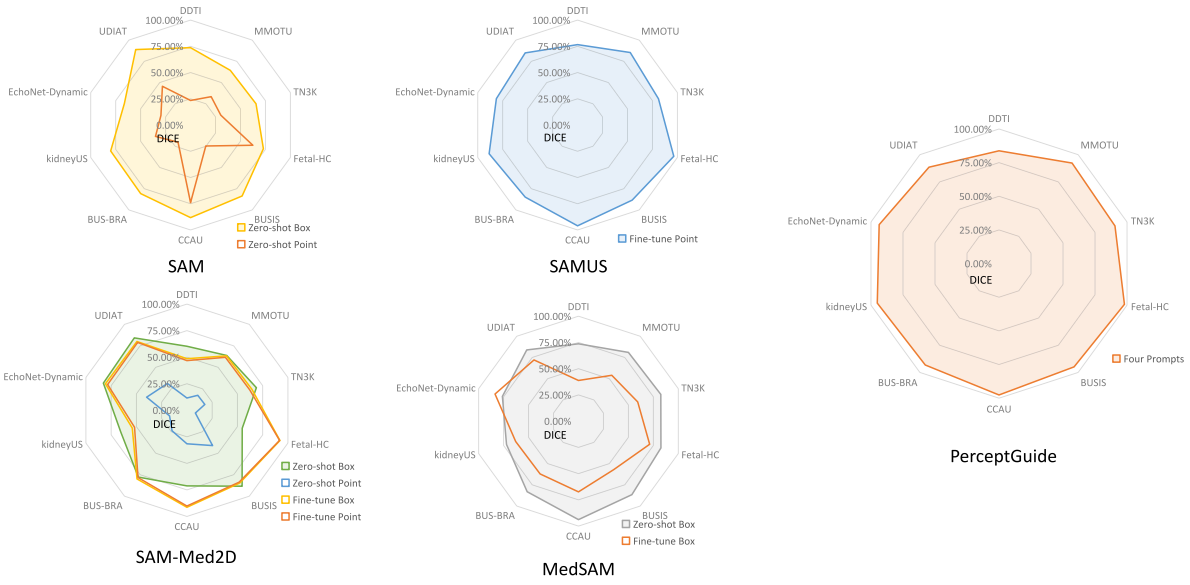

Fig. 6. Radar chart comparison of performance across different datasets for various models, including our proposed model and state-of-the-art (SOTA) models.

Table 5
Comparison of ablation study results across various datasets and tasks, highlighting the impact of hyper-perception and specialized models.

Dataset	Task	Specialist	Orchestration (with prompt guidance ?)	
			×	PerceptGuide (✓)
Params		29.59M/34.98M (505.78M)	66.13M	66.22M
DDTI	seg	64.16%	68.80%	**74.31%**
MMOTU	seg	75.68%	78.15%	**79.46%**
TN3K	seg	76.61%	74.97%	**77.76%**
Fetal-HC	seg	96.12%	96.53%	**96.74%**
BUSIS	seg	88.28%	91.15%	**91.56%**
CCAU	seg	92.97%	93.37%	**94.31%**
BUS-BRA	seg	84.00%	85.04%	**86.92%**
kidneyUS	seg	85.41%	85.62%	**90.39%**
EchoNet-Dynamic	seg	**91.80%**	91.16%	91.74%
UDIAT	seg	67.60%	72.54%	**81.32%**
seg		82.26%	83.73%	**86.45%**
TN3K	cls	64.66%	67.92%	**71.82%**
CUBS	cls	81.16%	78.99%	**86.23%**
BUS-BRA	cls	82.13%	84.00%	**88.27%**
Appendicitis	cls	**61.05%**	48.42%	46.32%
Fatty-Liver	cls	69.09%	86.36%	**90.91%**
UDIAT	cls	69.70%	81.82%	**90.91%**
cls		71.30%	74.58%	**79.08%**

and classification. From a parameter perspective, the specialist segmentation model is 29.59M, and the classification model is 34.98M. As previously noted, deploying all specialist models would require over 500M of loading space for 10 segmentation and 6 classification tasks. This many specialist models would also result in unaffordable running-time GPU RAM consumption. In contrast, the orchestration model only requires about 15% of the space to achieve even better average performance. The average performance of PerceptGuide surpasses both the specialist models and the results without prompt guidance, demonstrating that prompts significantly enhance network flexibility and prior knowledge injection. In summary, the integration of prompt prior knowledge and the flexibility provided by the hyper-perception module are key to PerceptGuide's superior performance over other methods.

To further dissect the contribution of the different prompt categories integrated via the hyper-perception module, we conducted a comprehensive component-wise ablation study. We evaluated all 16 combinations of the four prompt types: Position (P), Task (T), Input (I) and Object (O). The results, averaged across all segmentation and classification test sets, are presented in Table 6.

Table 6 reveals several key findings. Firstly, the full prompt combination (P✓T✓I✓O✓) yields the highest overall performance (83.69%), significantly improving upon the baseline ('0000', 80.30%). Secondly, removing any single prompt type reduces the total score, indicating a synergistic benefit from using all four prompts together. Thirdly, the impact varies by prompt type and task; for instance, the Position prompt ('P') appears particularly beneficial for segmentation, while achieving the best classification accuracy requires the full combination. This detailed analysis underscores the importance of integrating all four semantic prompt dimensions for optimal multi-task performance in PerceptGuide. We provide the extensive per-dataset results for all

Table 6
Component-wise ablation study of prompt contributions. Each binary entry indicates the presence (✓) or absence (0) of Position (P), Task (T), Input (I), and Object (O) prompts. Performance is averaged over all segmentation (Seg Dice) and classification (Cls Acc) datasets. 'Total' is the average of Seg and Cls scores.

Prompt configuration				Average performance		
P	T	I	O	Seg (Dice)	Cls (Acc)	Total
0	0	0	0	83.73%	74.58%	80.30%
0	0	0	✓	84.60%	66.78%	77.91%
0	0	✓	0	83.97%	75.46%	80.77%
0	0	✓	✓	82.76%	65.88%	76.43%
0	✓	0	0	81.97%	64.04%	75.25%
0	✓	0	✓	83.48%	71.90%	79.14%
0	✓	✓	0	83.30%	71.52%	78.89%
0	✓	✓	✓	85.94%	70.49%	80.15%
✓	0	0	0	84.94%	72.41%	80.24%
✓	0	0	✓	86.61%	72.75%	81.41%
✓	0	✓	0	84.19%	72.51%	79.81%
✓	0	✓	✓	**86.45%**	73.48%	81.59%
✓	✓	0	0	85.79%	74.39%	81.51%
✓	✓	0	✓	83.94%	73.65%	80.08%
✓	✓	✓	0	84.22%	73.46%	80.19%
✓	✓	✓	✓	**86.45%**	**79.08%**	**83.69%**

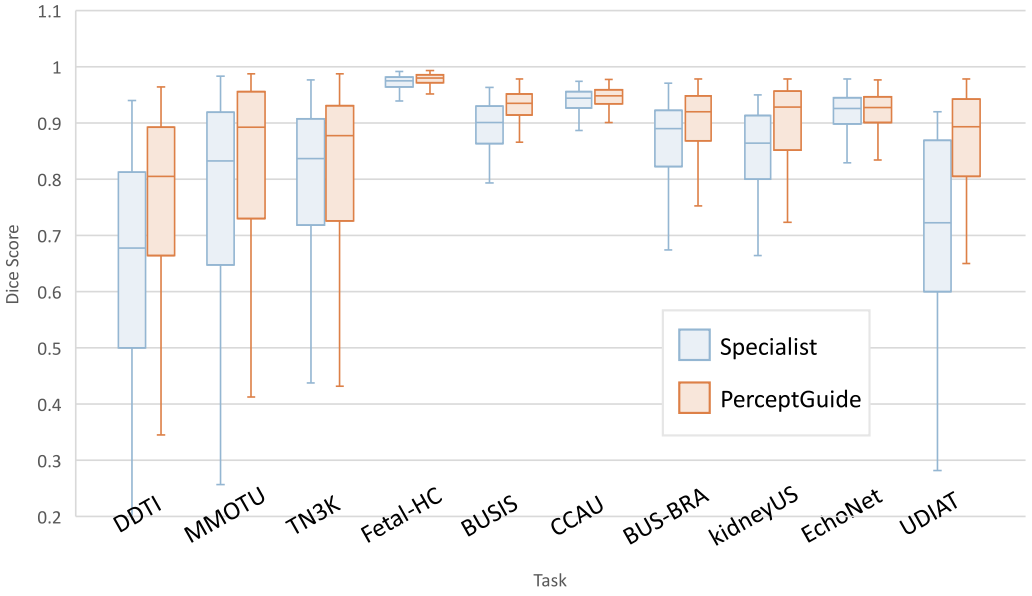

Fig. 7. Boxplot comparison of segmentation performance (Dice coefficient) between the Specialist (left) and PerceptGuide (right) models across various datasets.

16 configurations in our public code repository for interested readers (see Code and Data Availability Statement).

We utilize a boxplot to highlight the segmentation results of our model compared to the Specialist model across different datasets. The goal of developing a general-purpose model is to outperform specialist models in handling multiple datasets effectively and efficiently. As shown in Fig. 7, the qualitative visualization results indicate that our proposed general model, PerceptGuide, consistently surpasses the Specialist model across various datasets. This figure corresponds to Table 5, where we observed that PerceptGuide has superior average segmentation performance compared to the Specialist model (86.45% vs 82.26%). The boxplot results further confirm this from the perspective of median values, demonstrating that the performance distribution of our model is better than that of individual specialist models. These qualitative visualization results illustrate the versatility of our model in learning shared features across different ultrasound images, achieving strong performance across various datasets.

We also compared our model with the specialist models, an ablation model across different anatomical position. This figure complements Table 5, as shown in Fig. 8, our proposed model outperforms others in the majority of position, achieving the best performance in seven of the nine position (Head, Kidney, Liver, Thyroid, Breast, Ovary). In an additional position (Cardiac), the performance is nearly equivalent to the best results. The slightly lower performance in a position (Appendix) may be attributed to insufficient data, limiting feature learning. This chart demonstrates that our proposed model generally outperforms other models across different position, highlighting its strong generalization capability and ability to learn features from various anatomical locations.

4.3. External validation

The previous experimental results were based on internal validation using the M^2-US dataset, as shown in Fig. 5. Now, we present the external validation results using datasets introduced in Table 3. Two configurations are considered: Zero-Shot and Post-Training. Zero-Shot refers to directly testing pre-trained models on these external validation datasets. Due to specialist models being originally trained on specific datasets, we used models matching the position type for zero-shot testing; for example, the experiment on the BUSI dataset (Breast) used

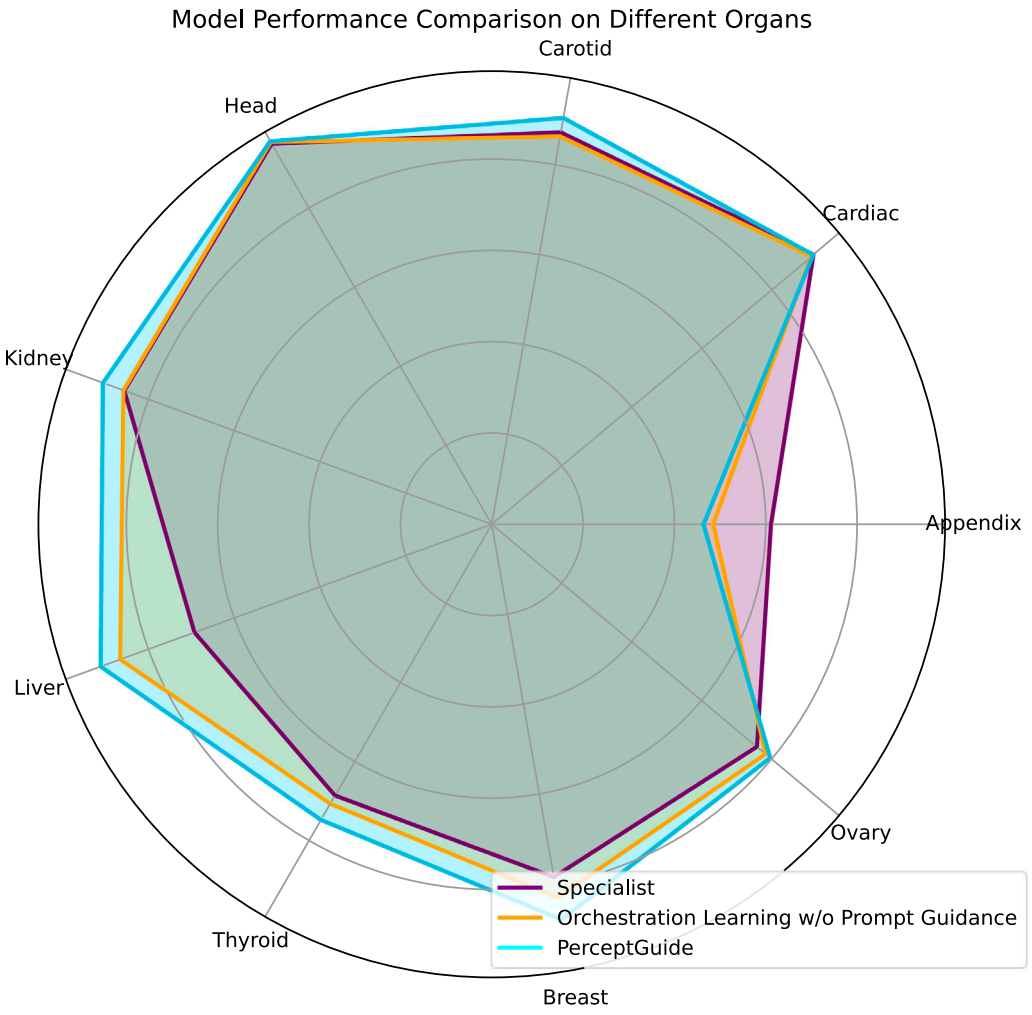

Fig. 8. Radar chart comparison of performance across different anatomical position for four models.

Table 7
External validation results for multiple datasets.

Dataset	Task	Zero-shot			Post-training	
		Specialist	Orchestration w/o Prompt	PerceptGuide	Specialist (Train for scratch)	PerceptGuide (Synchroniztion)
BUSI	cls	68.70%	66.41%	72.52%	63.36%	**75.57%**
BUSI	seg	62.04%	65.05%	67.66%	73.50%	**74.34%**
HMC-QU	seg	2.19%	5.88%	4.30%	**90.54%**	90.19%
TG3K	seg	0.41%	0.91%	0.76%	73.17%	**86.91%**
Average		33.34%	34.56%	36.31%	75.14%	**81.75%**

the pre-trained BUS-BRA model; the experiment on HMC-QU dataset (Cardiac) used EchoNet-Dynamic, and the experiment on TG3K dataset (Thyroid) used TN3K, matching the position type. For PerceptGuide and its ablation version, the trained models were used for zero-shot testing. In the Post-Training configuration, results were obtained by training a specialist model from scratch on each external dataset, and by fine-tuning our PerceptGuide model for a downstream synchronization stage. The table lists evaluation metrics, with segmentation using Dice and classification using accuracy. The best results for each dataset are bolded, and the best average results are underlined and bolded.

Analyzing the results, our model shows a clear advantage in zero-shot performance on the BUSI dataset (Breast). As shown in Table 7, the zero-shot results for HMC-QU (Cardiac) and TG3K (Thyroid) are unexpectedly low, possibly due to domain shifts or limited feature overlap with the training data. In the Post-Training comparison, our model shows clear improvements over specialist models trained from scratch, achieving competitive performance on the HMC-QU dataset (Cardiac). Despite conducting experiments on unseen datasets, our downstream synchronization stage benefits from prior exposure to images of the same regions during training, enabling efficient fine-tuning. In conclusion, the downstream synchronization stage enables cost-effective model transfer while maintaining high performance.

4.4. Insightful visualizations

We utilized Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to visualize the low-dimensional embeddings of nine different positions, with each position represented by

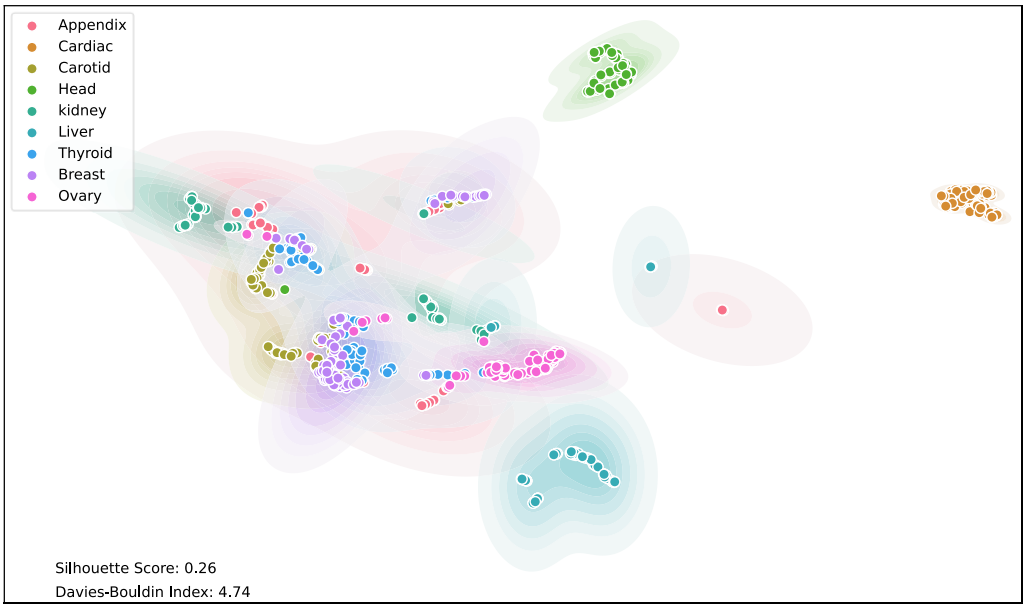

Fig. 9. UMAP visualization of low-dimensional embeddings for nine different positions.

a unique color. UMAP is a nonlinear dimensionality reduction and visualization algorithm. It leverages graph theory and manifold learning to map high-dimensional data into a lower-dimensional space for ease of visualization and analysis. Compared to other dimensionality reduction techniques such as *PCA* (Abdi and Williams, 2010) and *t-SNE* (Van der Maaten and Hinton, 2008), UMAP excels in preserving both the local and global structure of the data, making it effective for handling large-scale datasets. We extracted features from the encoder of various position images and applied UMAP to these features, allowing us to observe relationships between different positions in a two-dimensional space. As shown in Fig. 9, the UMAP projection displays discernible clustering, with some organs like Cardiac, Head, and Liver forming relatively distinct groups, while others exhibit overlap, suggesting the model learns both position-specific and potentially shared cross-organ features. This visual impression is corroborated by quantitative metrics (Silhouette Score: 0.26, Davies–Bouldin Index: 4.74) calculated on the pre-projection embeddings, indicating moderate overall cluster separation consistent with the observed patterns.

We present the visual results of our model's segmentation and classification tasks. As shown in Fig. 10, the first row displays the original ultrasound images. The subsequent four rows represent the prompts input to our model, corresponding to each image in the first row. The final row illustrates the outcomes of either segmentation or classification.

The qualitative visualization results demonstrate that our model achieves satisfactory performance across a range of anatomical position in ultrasound images. The incorporation of prior knowledge through Position prompts enhances the model's performance in segmenting and classifying ultrasound images from various position. Task prompts allow the model to adapt its inductive bias between segmentation and classification tasks, enhancing generalization capabilities. Object prompts enhance adaptability, allowing the model to manage heterogeneity and homogeneity across different object types. Input prompts facilitate processing images at varying magnifications, establishing a spatial granularity guideline for the model. The *Highlighted* Input prompts, in particular, maximize localization information to assist classification. In practical applications, initial segmentation can be conveniently provided by clinicians or generated by segmentation models. These qualitative visualization results indicate that our model exhibits strong generalization and adaptability in segmentation and classification tasks across various ultrasound image position.

4.5. Qualitative results and failure case analysis

While PerceptGuide demonstrates strong overall performance across various tasks and datasets, analyzing specific instances where the model underperforms provides valuable insights into its current limitations and potential areas for improvement. Fig. 11 presents representative examples of such failure cases encountered during our evaluation.

The top row (Fig. 11a-c) showcases segmentation challenges. In Fig. 11a, the segmentation of an ovarian tumor is incomplete, possibly due to complex internal structures or indistinct boundaries often seen in such lesions. Fig. 11b illustrates a breast nodule segmentation where the prediction (yellow contour) appears significantly influenced by artifacts in the nearby glandular tissue, leading to an inaccurate delineation extending beyond the actual nodule (green contour). For the thyroid nodule in Fig. 11c, the model struggles with the highly ambiguous boundary between the nodule and the surrounding parenchyma, incorrectly segmenting parts of the healthy gland instead of adhering closely to the subtle nodule edges indicated by the ground truth.

The bottom row (Fig. 11d-f) highlights classification difficulties. Fig. 11d shows a liver image where the presence of significant image noise and texture variations might have contributed to the incorrect classification regarding the presence of fatty liver. In Fig. 11e, the model misclassifies a breast nodule potentially due to its atypical presentation; while posterior acoustic shadowing (often indicative of malignancy) is present, the model may have over-relied on this single feature without adequately integrating other potentially benign indicators, leading to an incorrect assessment compared to the ground truth label. Finally, Fig. 11f presents a thyroid nodule classification failure likely caused by the very low contrast between the nodule and the adjacent thyroid tissue, making the grayscale distinction inherently difficult and challenging the model's feature extraction capabilities.

In summary, these failure cases underscore common challenges in ultrasound image analysis that can affect automated models, including: (i) image quality issues such as noise, artifacts, and low contrast; (ii) inherent ambiguity in lesion boundaries or tissue differentiation; and (iii) atypical presentations of disease that deviate significantly

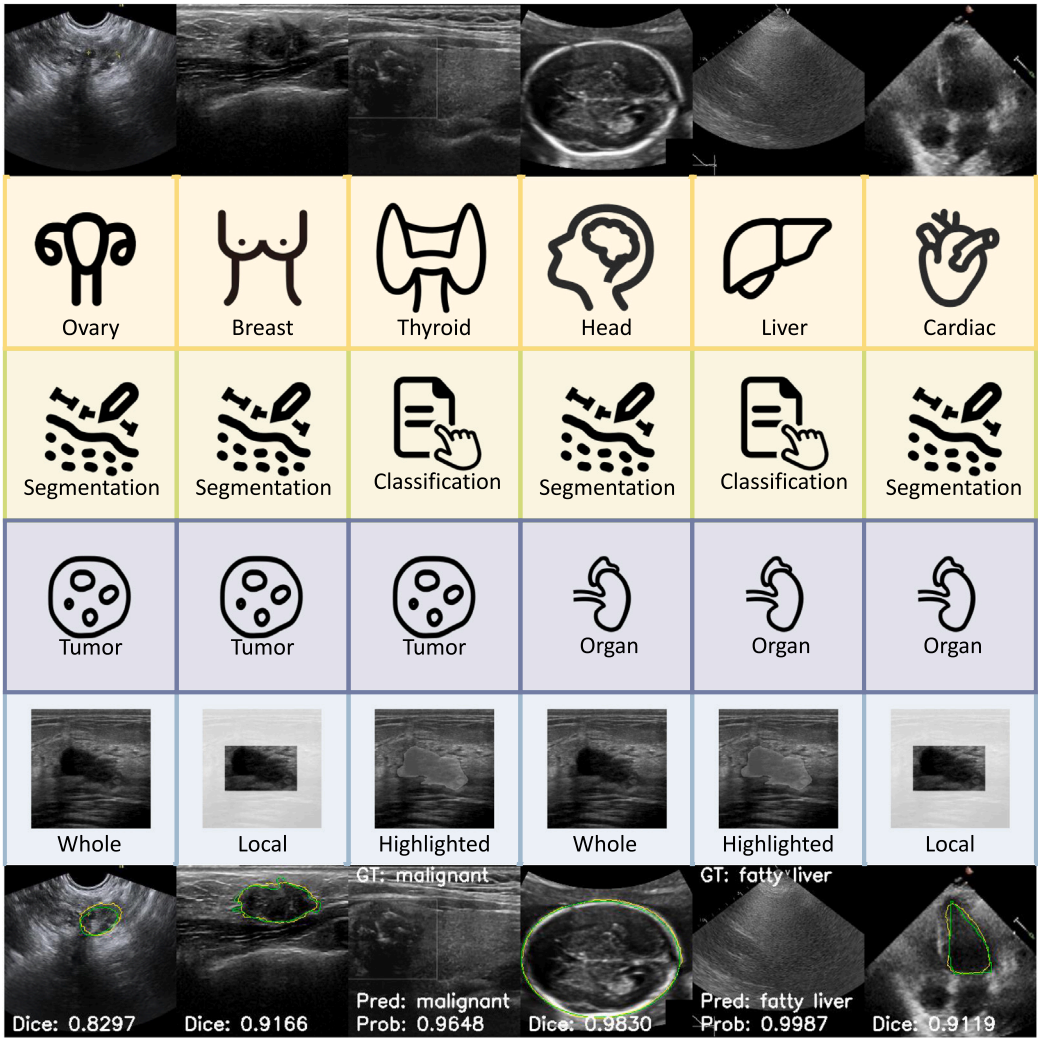

Fig. 10. Qualitative results of ultrasound image segmentation and classification using various prompts.

from learned patterns. Addressing these challenges, potentially through improved pre-processing techniques, advanced artifact handling, the integration of more sophisticated reasoning mechanisms, or training on even larger and more diverse datasets encompassing rare cases, remains an important direction for future research to enhance the robustness and clinical applicability of automated ultrasound analysis frameworks like PerceptGuide.

4.6. Clinical applications and future directions

Our model's effectiveness is primarily attributed to two key factors: the integration of prior knowledge and the enhancement of network plasticity and flexibility. These elements allow the model to adapt to diverse clinical scenarios, making it a valuable tool in the medical field. Our experiments show that PerceptGuide achieves a fast inference speed (approximately **37–38** FPS on the BUS-BRA test set) using an NVIDIA RTX A4000 GPU (16GB). Furthermore, the model requires only modest GPU memory (**max 1.4GB, mean 1.0GB** during inference). This demonstrates the model's potential for near real-time application and its feasibility on common clinical workstations. We contrast this favorably with the significantly higher cumulative memory footprint and potential loading times associated with deploying numerous individual specialist models (as indicated by the parameter counts in Table 5). We envision that these contextual prompts could be automatically selected based on the type of examination being performed. Alternatively, a lightweight helper module could predict appropriate prompts, or the interface could allow optional user selection/confirmation, offering flexibility while maintaining the core automated capability.

However, certain limitations exist. For instance, the Position option "Other" lacks specialized training, presenting an opportunity for future investigation from an out-of-distribution (OOD) perspective. Addressing this could improve the model's robustness and adaptability to unforeseen categories. Currently, our model has been validated on publicly available datasets. Future efforts should focus on expanding validation to a broader array of datasets, including collaborations with healthcare institutions for clinical trials. This would ensure the model's reliability and effectiveness in real-world settings.

Another critical aspect to consider is the evolving nature of medical image processing itself. While conventional approaches often treat tasks such as segmentation, classification, and retrieval as isolated perceptual tasks, we argue that medical imaging is better understood as a chain-of-thought (CoT) process. This process requires breaking down complex imaging problems into multiple functional modules and leveraging multi-hop reasoning to arrive at accurate conclusions. This perspective aligns with the increasing recognition that medical image analysis must incorporate higher-order reasoning and decision-making steps to match clinical expectations.

Additionally, the choice of training data plays a pivotal role in the robustness and generalizability of AI models. Relying solely on real-world hospital reports for training is often suboptimal due to noisy and inconsistent annotations, as well as the high marginal cost

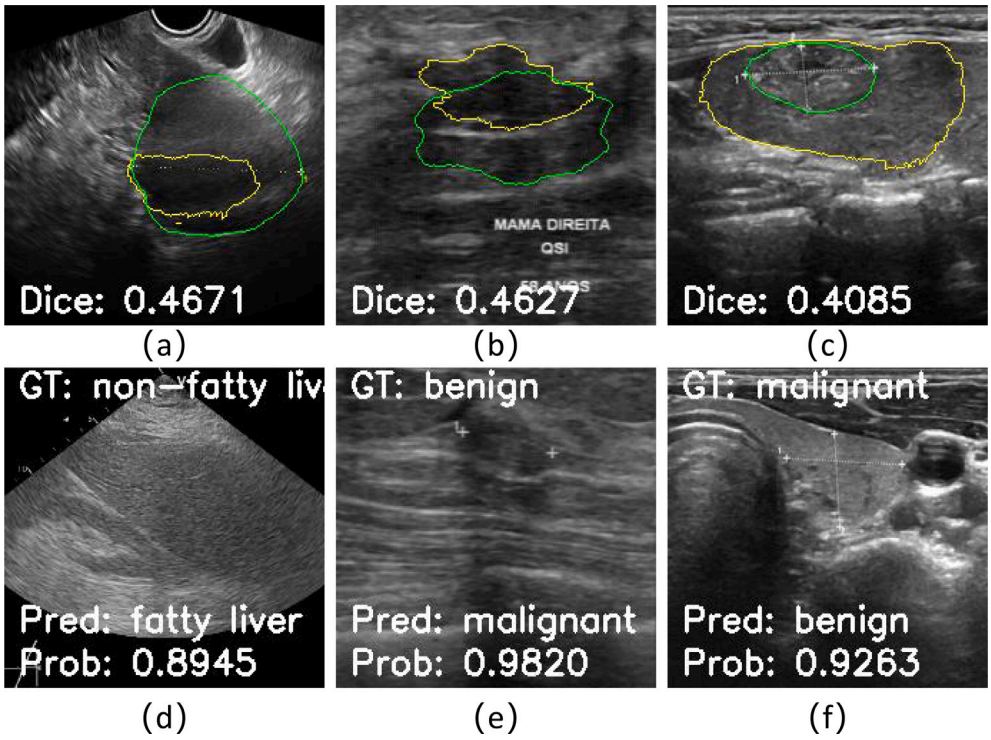

Fig. 11. Examples of Failure Cases for PerceptGuide. Rows 1 and 2 show representative examples of segmentation and classification failures, respectively. (a-c) Segmentation examples for (a) Ovarian tumor, (b) Breast nodule, and (c) Thyroid nodule. Green contours represent ground truth (GT), yellow contours represent model predictions. Dice scores are provided below each image. (d-f) Classification examples for (d) Fatty liver detection, (e) Benign/malignant breast nodule classification, and (f) Benign/malignant thyroid nodule classification. GT labels are shown above, predicted labels and probabilities are shown below.

of data curation. Instead, structured and high-quality datasets—such as those derived from textbooks and internet case studies—offer a promising alternative. These datasets, being richly annotated and free from the inconsistencies found in real-world clinical reports, could significantly improve model performance while reducing the cost of data preparation. Exploring such high-quality data sources for pre-training could also pave the way for better generalization to noisy, real-world environments.

Furthermore, there is potential to enhance the model's zero-shot capabilities. Future research could explore the integration of self-supervised learning techniques, leveraging pre-training with radiology reports to strengthen the model's ability to generalize without extensive organ-specific data. Further research could also explore the incorporation of multi-modal data, combining imaging with clinical metadata to enrich the model's predictive power. This holistic approach could lead to more personalized and accurate diagnostic tools. Moreover, investigating the ethical implications and ensuring the transparency of AI models in clinical settings will be crucial. Establishing clear guidelines and maintaining openness in AI-driven decisions will foster trust and facilitate widespread adoption in healthcare.

5. Conclusion

In this study, we presented PerceptGuide, an orchestration learning framework for ultrasound image classification and segmentation. By integrating the hyper-perception module and leveraging four specially designed prompts (Object, Task, Input, and Position) for ultrasound, our approach enhances adaptability across multiple datasets and organs. Evaluated on the comprehensive M^2-US dataset, PerceptGuide demonstrates robust performance, surpassing existing AI solutions. The introduction of a downstream synchronization training stage further improves generalization capabilities. This work offering efficient diagnostic support in clinical practice. Future work will focus on expanding the dataset and refining the framework to further improve its applicability and performance in real-world clinical settings.

Code and data availability statement

The M^2-US dataset utilized in this study is a compilation derived from 16 publicly available datasets. Detailed instructions for obtaining the source datasets, the scripts used for pre-processing and compiling M^2-US, along with the source code for the PerceptGuide framework and pre-trained model weights, are publicly available on GitHub at: https://github.com/Zehui-Lin/PerceptGuide. Access procedures vary for the original datasets based on their respective licenses, as detailed in the repository's documentation.

CRediT authorship contribution statement

Zehui Lin: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Shuo Li:** Writing – review & editing, Formal analysis, Conceptualization. **Shanshan Wang:** Writing – review & editing, Formal analysis, Conceptualization. **Zhifan Gao:** Writing – review & editing, Formal analysis, Conceptualization. **Yue Sun:** Writing – review & editing, Supervision, Project administration, Formal analysis. **Chan-Tong Lam:** Project administration, Funding acquisition, Conceptualization. **Xindi Hu:** Resources, Formal analysis, Conceptualization. **Xin Yang:** Resources, Formal analysis, Conceptualization. **Dong Ni:** Resources, Formal analysis, Conceptualization. **Tao Tan:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Science and Technology Development Fund of Macao (0021/2022/AGJ) and Science and Technology Development Fund of Macao (0041/2023/RIB2).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2025.103639.

Data availability

The authors do not have permission to share data.

References

Abdi, H., Williams, L.J., 2010. Principal component analysis. Wiley Interdiscip. Rev.: Comput. Stat. 2 (4), 433–459.

Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A., 2020. Dataset of breast ultrasound images. Data Brief 28, 104863.

Byra, M., Styczynski, G., Szmigielski, C., Kalinowski, P., Michałowski, Ł., Paluszkiewicz, R., Ziarkiewicz-Wróblewska, B., Zieniewicz, K., Sobieraj, P., Nowicki, A., 2018. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. Int. J. Comput. Assist. Radiol. Surg. 13, 1895–1903.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. Springer, pp. 205–218.

Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al., 2023. Sam-med2d. arXiv preprint arXiv:2308.16184.

Choe, J., Shim, H., 2019. Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2219–2228.

Degerli, A., Zabihi, M., Kiranyaz, S., Hamid, T., Mazhar, R., Hamila, R., Gabbouj, M., 2021. Early detection of myocardial infarction in low-quality echocardiography. IEEE Access 9, 34442–34453.

Ding, K., Zhou, M., Wang, H., Gevaert, O., Metaxas, D., Zhang, S., 2023. A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. Sci. Data 10 (1), 231.

Ferraioli, G., Monteiro, L.B.S., 2019. Ultrasound-based techniques for the diagnosis of liver steatosis. World J. Gastroenterol. 25 (40), 6053.

Folland, E., Parisi, A., Moynihan, P., Jones, D.R., Feldman, C.L., Tow, D., 1979. Assessment of left ventricular ejection fraction and volumes by real-time, two-dimensional echocardiography. A comparison of cineangiographic and radionuclide techniques. Circulation 60 (4), 760–766.

Gómez-Flores, W., Gregorio-Calas, M.J., Coelho de Albuquerque Pereira, W., 2024. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. Med. Phys. 51 (4), 3110–3123. http://dx.doi.org/10.1002/mp.16812, arXiv:https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.16812 URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.16812.

Gong, H., Chen, J., Chen, G., Li, H., Li, G., Chen, F., 2023. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. Comput. Biol. Med. 155, 106389.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009.

He, S., Lin, Z., Yang, X., Chen, C., Wang, J., Shuang, X., Deng, Z., Liu, Q., Cao, Y., Lu, X., et al., 2021. Statistical dependency guided contrastive learning for multiple labeling in prenatal ultrasound. In: Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12. Springer, pp. 190–198.

He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H., 2023. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Trans. Med. Imaging 42 (9), 2763–2775.

Hua, S., Yan, F., Shen, T., Zhang, X., 2023. Pathoduet: Foundation models for pathological slide analysis of h&e and ihc stains. arXiv preprint arXiv:2312.09894.

Huang, Z., Deng, Z., Ye, J., Wang, H., Su, Y., Li, T., Sun, H., Cheng, J., Chen, J., He, J., et al., 2023a. A-Eval: A benchmark for cross-dataset evaluation of abdominal multi-organ segmentation. arXiv preprint arXiv:2309.03906.

Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al., 2023b. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18 (2), 203–211.

Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., et al., 2024. USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Med. Image Anal. 96, 103202.

Kang, Q., Gao, J., Li, K., Lao, Q., 2023. Deblurring masked autoencoder is better recipe for ultrasound image recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 352–362.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026.

Lei, W., Wei, X., Zhang, X., Li, K., Zhang, S., 2023. Medlsam: Localize and segment anything model for 3d medical images. arXiv preprint arXiv:2306.14752.

Li, Z., Shang, Z., Liu, J., Zhen, H., Zhu, E., Zhong, S., Sturgess, R.N., Zhou, Y., Hu, X., Zhao, X., et al., 2023. D-LMBmap: a fully automated deep-learning pipeline for whole-brain profiling of neural circuitry. Nature Methods 20 (10), 1593–1604.

Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al., 2022. Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975.

Lin, X., Xiang, Y., Zhang, L., Yang, X., Yan, Z., Yu, L., 2023. SAMUS: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. arXiv preprint arXiv:2309.06824.

Lin, Z., Zhang, Z., Hu, X., Gao, Z., Yang, X., Sun, Y., Ni, D., Tan, T., 2024. UniUSNet: A promptable framework for universal ultrasound disease prediction and tissue segmentation. In: 2024 IEEE International Conference on Bioinformatics and Biomedicine. BIBM, IEEE, pp. 3501–3504.

Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B., 2024. Segment anything in medical images. Nat. Commun. 15 (1), 654.

Marcinkevičs, R., Reis Wolfertstetter, P., Klimiene, U., Chin-Cheong, K., Paschke, A., Zerres, J., Denzinger, M., Niederberger, D., Wellmann, S., Ozkan, E., Knorr, C., Vogt, J.E., 2023. Regensburg Pediatric Appendicitis Dataset. Zenodo, http://dx.doi.org/10.5281/zenodo.7711412.

McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Meiburger, K.M., Marzola, F., Zahnd, G., Faita, F., Loizou, C.P., Lainé, N., Carvalho, C., Steinman, D.A., Gibello, L., Bruno, R.M., et al., 2022. Carotid Ultrasound Boundary Study (CUBS): Technical considerations on an open multi-center analysis of computerized measurement systems for intima-media thickness measurement on common carotid artery longitudinal B-mode ultrasound scans. Comput. Biol. Med. 144, 105333.

Momot, A., 2022. Common carotid artery ultrasound images. Mendeley Data, http://dx.doi.org/10.17632/d4xt63mgjm.1.

Mostbeck, G., Adam, E.J., Nielsen, M.B., Claudon, M., Clevert, D., Nicolau, C., Nyhsen, C., Owens, C.M., 2016. How to diagnose acute appendicitis: ultrasound first. Insights Into Imaging 7, 255–263.

Mostbeck, G.H., Zontsich, T., Turetschek, K., 2001. Ultrasound of the kidney: obstruction and medical diseases. Eur. Radiol. 11, 1878–1889.

Mullapudi, R.T., Mark, W.R., Shazeer, N., Fatahalian, K., 2018. Hydranets: Specialized dynamic architectures for efficient inference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8080–8089.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al., 2020. Video-based AI for beat-to-beat assessment of cardiac function. Nature 580 (7802), 252–256.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E., 2015. An open access thyroid ultrasound image database. In: 10th International Symposium on Medical Information Processing and Analysis, vol. 9287, SPIE, pp. 188–193.

Qin, Z., Yi, H., Lao, Q., Li, K., 2022. Medical image understanding with pretrained vision language models: A comprehensive study. arXiv preprint arXiv:2209.15517.

Singla, R., Ringstrom, C., Hu, G., Lessoway, V., Reid, J., Nguan, C., Rohling, R., 2023. The open kidney ultrasound data set. In: International Workshop on Advances in Simplifying Medical Ultrasound. Springer, pp. 155–164.

Spak, D.A., Plaxco, J., Santiago, L., Dryden, M., Dogan, B., 2017. BI-RADS® fifth edition: A summary of changes. Diagn. Interv. Imaging 98 (3), 179–190.

Stein, J.H., Korcarz, C.E., Hurst, R.T., Lonn, E., Kendall, C.B., Mohler, E.R., Najjar, S.S., Rembold, C.M., Post, W.S., 2008. Use of carotid ultrasound to identify subclinical vascular disease and evaluate cardiovascular disease risk: a consensus statement from the American Society of Echocardiography Carotid Intima-Media Thickness Task Force endorsed by the Society for Vascular Medicine. J. Am. Soc. Echocardiogr. 21 (2), 93–111.

Sun, S., Fu, C., Xu, S., Wen, Y., Ma, T., 2024. CRSANet: Class representations self-attention network for the segmentation of thyroid nodules. Biomed. Signal Process. Control. 91, 105917.

Tessler, F.N., Middleton, W.D., Grant, E.G., Hoang, J.K., Berland, L.L., Teefey, S.A., Cronan, J.J., Beland, M.D., Desser, T.S., Frates, M.C., et al., 2017. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. J. Am. Coll. Radiol. 14 (5), 587–595.

van den Heuvel, T.L.A., de Bruijn, D., de Korte, C.L., van Ginneken, B., 2018. Automated Measurement of Fetal Head Circumference Using 2D Ultrasound Images. Zenodo, http://dx.doi.org/10.5281/zenodo.1327317.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y., 2023a. SAM-Med3D. arXiv:2310.15161.

Wang, Z., Liu, C., Zhang, S., Dou, Q., 2023b. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 101–111.

Wang, D., Wang, X., Wang, L., Li, M., Da, Q., Liu, X., Gao, X., Shen, J., He, J., Shen, T., et al., 2023c. A real-world dataset and benchmark for foundation model adaptation in medical image classification. Sci. Data 10 (1), 574.

Wang, G., Wu, J., Luo, X., Liu, X., Li, K., Zhang, S., 2023d. Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset. arXiv preprint arXiv:2306.16925.

Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z., 2022. FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. Med. Image Anal. 76, 102327.

Wu, Y., Li, S., Du, Z., Zhu, W., 2023. BROW: Better features for whole slide image based on self-distillation. arXiv preprint arXiv:2309.08259.

Wu, C., Wang, Y., Wang, F., 2018. Deep learning for ovarian tumor classification with ultrasound images. In: Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part III 19. Springer, pp. 395–406.

Wunderling, T., Golla, B., Poudel, P., Arens, C., Friebe, M., Hansen, C., 2017. Comparison of thyroid segmentation techniques for 3D ultrasound. In: Medical Imaging 2017: Image Processing, vol. 10133, SPIE, pp. 346–352.

Yang, H., Yang, D., 2023. CSwin-PNet: A CNN-Swin Transformer combined pyramid network for breast lesion segmentation in ultrasound images. Expert Syst. Appl. 213, 119024.

Yap, M.H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., Davison, A.K., Marti, R., 2017. Automated breast ultrasound lesions detection using convolutional neural networks. IEEE J. Biomed. Heal. Informatics 22 (4), 1218–1226.

Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D., 2023. Text-guided foundation model adaptation for pathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 272–282.

Zhang, Y., Liu, H., Hu, Q., 2021. Transfuse: Fusing transformers and cnns for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, pp. 14–24.

Zhang, S., Metaxas, D., 2023. On the challenges and perspectives of foundation models for medical image analysis. Med. Image Anal. 102996.

Zhang, Y., Xian, M., Cheng, H.-D., Shareef, B., Ding, J., Xu, F., Huang, K., Zhang, B., Ning, C., Wang, Y., 2022. BUSIS: a benchmark for breast ultrasound image segmentation. In: Healthcare, vol. 10, (4), MDPI, p. 729.

Zhao, Q., Lyu, S., Bai, W., Cai, L., Liu, B., Wu, M., Sang, X., Yang, M., Chen, L., 2022. A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. arXiv preprint arXiv:2207.06799.

Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W., 2023. One model to rule them all: Towards universal segmentation for medical images with text prompts. arXiv preprint arXiv:2312.17183.

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