FreqDINO++: A Frequency-Guided Multi-Task Routing Vision Foundation Model for Universal Ultrasound Analysis

Qing Xu, Yixuan Zhang, Yue Li, Xiangjian He, *Senior Member, IEEE*, Qian Zhang, Mainul Haque, Rong Qu, *Fellow, IEEE*, Wenting Duan, Jieyun Bai, Zhen Chen

***Abstract*— Ultrasound image analysis plays a crucial role in cancer screening and prenatal diagnosis, yet comprehensive assessment requires jointly addressing tasks such as lesion segmentation and benign-malignant classification. While recent vision foundation models have shown remarkable universal representations, unlocking their potential for ultrasound is bottlenecked by the considerable domain gap from natural images. Existing methods typically fine-tune heavy vision encoders for isolated tasks, incurring substantial computational overhead while overlooking the underlying commonalities across heterogeneous tasks. In this work, we propose FreqDINO++, a frequency-guided multi-task routing vision foundation model for universal ultrasound analysis. We first introduce a Multi-task Routing Adapter (MR-Adapter) to support parameter-efficient integration of task-common and task-specific knowledge, a Frequency-aware Feature Enhancer (F^2-Enhancer) is then designed to capture the rich multi-scale frequency characteristics of ultrasound images, and a Task-aligned Collaborative Decoder (TC-Decoder) is devised to promote collaboration between dense and global prediction tasks through global-local token interaction. Extensive experiments on large-scale multi-task and external single-task ultrasound benchmarks demonstrate that FreqDINO++ consistently outperforms strong baselines and recent foundation models across 27 diverse clinical task scenarios, while also showing promising generalization to unseen data. The code is at https://github.com/MingLang-FD/FreqDINO-Plus.**

***Index Terms*— Ultrasound analysis, foundation model, multi-task learning, frequency-guided adaptation**

This work is partially supported by the Zhejiang Department of Transportation General Research and Development Project (2024039), National Natural Science Foundation of China grant (62476037), and Ningbo Science and Technology Bureau under NBNSF General Programme (2025J114).*(Equal contribution: Q. Xu, Y. Zhang and L. Yue, Corresponding author: X. He and Z. Chen)*

Q. Xu, Y. Zhang, L. Yue, X. He, and Q. Zhang are with School of Computer Science, University of Nottingham Ningbo China, Ningbo, Zhejiang, China (e-mail: sean.he@nottingham.edu.cn).

M. Haque is with School of Mathematical Sciences, University of Nottingham Ningbo China, Ningbo, Zhejiang, China (e-mail: Mainul.Haque@nottingham.edu.cn).

R. Qu is with the School of Computer Science, University of Nottingham, Nottingham NG72RD, UK (email: rong.qu@nottingham.ac.uk).

W. Duan is with School of Engineering and Physical Science, University of Lincoln, Lincoln LN6 7TS, UK (email: wduan@lincoln.ac.uk).

J. Bai is with College of Information Science and Technology, Jinan University, Guangzhou, China (email: jbai996@aucklanduni.ac.nz).

Z. Chen is with Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong SAR. (e-mail: z.chen@polyu.edu.hk).

I. Introduction

Ultrasound image analysis plays a crucial role in clinical practice and has received significant research attention owing to its real-time, non-invasive, and cost-effective imaging characteristics [1]–[3]. For example, breast ultrasound examination requires precise lesion segmentation for volumetric estimation, diagnostic classification for benign-malignant differentiation, and bounding box localization for multi-lesion detection [4]–[6]. Similarly, obstetric ultrasound relies on anatomical structure segmentation for organ delineation, fetal standard plane classification for quality control, and keypoint regression for biometric measurements such as femur length and angle of progression [7], [8]. These demands of heterogeneous tasks have given rise to the challenging field of universal ultrasound image analysis, which seeks to unify segmentation, classification, detection, and regression within a single framework to serve diverse clinical workflows.

The classical ultrasound image analysis methods [2], [4], [9] have been designed for specific tasks, resulting in isolated solutions that address segmentation, classification, detection, and regression separately, as shown in Fig. 1(a). These conventional approaches typically adopt independent architectures without shared representations during training. For instance, segmentation networks [10]–[14] are optimized to delineate tumor or organ boundaries at the pixel level for volumetric quantification, classification models [15]–[17] differentiate benign-malignant lesions or recognize fetal standard planes for diagnostic support, detection frameworks [5], [6], [18] localize multiple lesions with bounding boxes for early screening, and regression methods [8], [19] estimate anatomical landmarks such as head circumference or femur length for biometric assessment. However, maintaining multiple task-specific networks [20], [21] leads to substantial computational overhead and deployment burden in clinical settings, motivating the development of a unified framework with shared representations that could serve multiple ultrasound analysis tasks simultaneously.

To this end, Vision Foundation Models (VFMs) pre-trained on large-scale datasets have emerged as a promising backbone for ultrasound image analysis. In particular, the recent DINOv3 [22] has demonstrated remarkable representation capacity through self-supervised learning, offering high-quality features

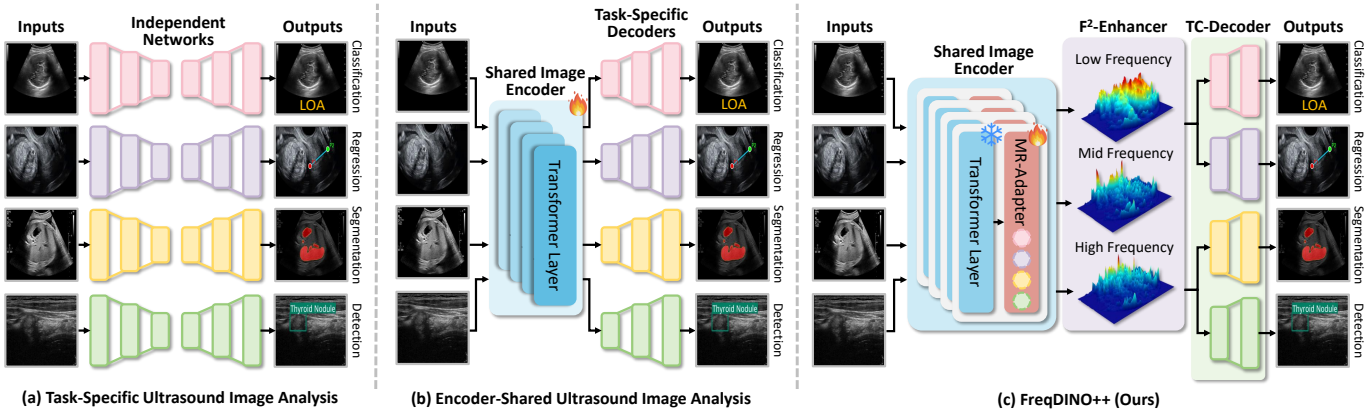

Fig. 1. Comparison of our FreqDINO++ and existing ultrasound image analysis works. (a) Independent task-specific networks separately handle heterogeneous ultrasound tasks. (b) A shared vision encoder with task-specific decoders for heterogeneous ultrasound tasks. (c) Our FreqDINO++ leverages multi-task routing and frequency-guided enhancement with token-collaborative decoding for universal ultrasound analysis.

for diverse downstream tasks. However, existing methods [4], [23], [24] typically rely on fully fine-tuning the heavy pre-trained encoder for individual task, which is computationally costly and hinders multi-task deployment in clinical practice, as shown in Fig. 1(b). Moreover, these methods typically treat each task in isolation and underexploit the underlying commonalities across heterogeneous tasks, where segmentation and detection share dense prediction cues such as boundaries and spatial localization, while classification and regression rely on similar global semantic context. More critically, because DINOv3 is pre-trained on natural images, its representations do not explicitly account for the low contrast, speckle noise, and frequency-dependent textures characteristic of ultrasound images [24]–[26]. These observations motivate us to develop a unified framework that efficiently adapts VFM to ultrasound imaging while coordinating heterogeneous downstream tasks for comprehensive clinical analysis.

To overcome these limitations, we propose FreqDINO++, a frequency-guided multi-task routing vision foundation model for universal ultrasound analysis. As illustrated in Fig. 1(c), FreqDINO++ adapts the frozen DINOv3 backbone to ultrasound scenarios through task-conditioned routing and frequency-guided enhancement within a unified framework. Specifically, we devise a Multi-Task Routing Adapter (MR-Adapter) to seamlessly unify task-common and task-specific knowledge for parameter-efficient multi-task adaptation. We further design a Frequency-aware Feature Enhancer (F^2-Enhancer) to fully exploit the multi-scale frequency characteristics of ultrasound images for discriminative feature representations. Moreover, we devise a Task-aligned Collaborative Decoder (TC-Decoder) that harnesses the interaction between global and local token representations to coordinate heterogeneous downstream tasks. Extensive experiments on diverse ultrasound datasets demonstrate that FreqDINO++ consistently outperforms state-of-the-art methods across all task categories.

The contributions of this work are summarized as follows:

- We propose FreqDINO++, a frequency-guided multi-task routing vision foundation model that adapts DINOv3 to universal ultrasound analysis, unifying heterogeneous clinical tasks within a single framework.
- We design the MR-Adapter coupled with F^2-Enhancer that seamlessly unifies task-common and task-specific knowledge through task-conditioned routing, while fully exploiting the multi-scale frequency characteristics inherent in ultrasound images for discriminative feature representations.
- We devise the TC-Decoder that harnesses the collaborative interaction between global and local token representations to coordinate heterogeneous downstream tasks, facilitating complementary knowledge transfer across dense and global prediction objectives.
- We validate FreqDINO++ on large-scale multi-task and external benchmark datasets, where it demonstrates strong performance across 27 clinical task scenarios and promising generalization to unseen data.

A preliminary version of this work has been published in ISBI 2026 [27]. In this paper, we have made substantial extensions with the following highlights: 1) We extend the scope from single-task segmentation to universal ultrasound analysis that jointly addresses segmentation, classification, detection, and regression; 2) We devise the MR-Adapter that replaces the generic lightweight adapter in the conference work [27], integrating task-common and task-specific knowledge through task-conditioned routing; 3) We reformulate the segmentation-oriented MFEA and FGBR into a unified F^2-Enhancer, extending frequency modeling from one segmentation task to 27 diverse clinical scenarios with multi-scale frequency characteristics; 4) We upgrade the dual-head boundary-guided decoder to a TC-Decoder that coordinates heterogeneous tasks via global-local token interaction; 5) We conduct extensive experiments on five datasets including the large-scale FMC-UIA challenge spanning 27 clinical task scenarios, together with comprehensive ablation studies.

II. Related Work

A. Ultrasound Image Analysis

Ultrasound imaging has become one of the most widely used modalities in clinical practice, owing to its real-time capability, non-invasive nature, and cost-effectiveness [1], [3]. In routine clinical workflows, ultrasound examination

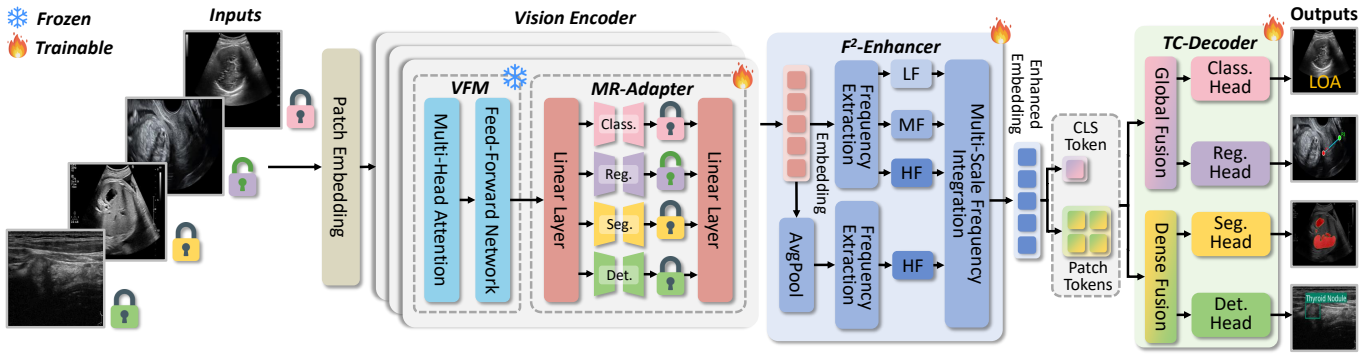

Fig. 2. The overview of our FreqDINO++ framework for universal ultrasound analysis, consisting of an MR-Adapter that integrates task-common and task-specific knowledge through task-conditioned routing, an F²-Enhancer that captures multi-scale frequency characteristics inherent in ultrasound images, and a TC-Decoder with global and dense heads for coordinating heterogeneous downstream tasks. The FreqDINO++ framework efficiently adapts the VFM to ultrasound scenarios and coordinates heterogeneous tasks within a unified architecture.

often requires the joint execution of multiple analytical tasks to support comprehensive patient assessment. For instance, breast ultrasound relies on precise lesion segmentation for volumetric estimation, diagnostic classification for benign-malignant differentiation, and bounding box localization for multi-lesion detection [4], [5]. Similarly, obstetric ultrasound demands anatomical structure segmentation for organ delineation, standard plane classification for quality control, and keypoint regression for biometric measurements [7], [8]. However, ultrasound images inherently suffer from speckle noise, low contrast, and frequency-dependent texture variations [9], [24], posing significant challenges for automated analysis. These heterogeneous task demands and imaging characteristics have motivated the development of advanced computational methods for universal ultrasound image analysis.

In this context, deep learning methods have made notable strides across individual ultrasound analysis tasks through tailored network architectures. For segmentation, U-Net [10] and its variants [11], [12], [21] have served as the predominant architectures, with subsequent works introducing attention mechanisms [4], [20], transformer-based designs [23], and frequency-domain fusion strategies [9]. For classification and detection, vision transformers [16], [17] and DETR-based frameworks [5], [6], [18] have been adopted for diagnostic categorization and lesion localization, respectively. Despite the progress, these methods are predominantly designed for specific tasks in isolation, requiring separate optimization and deployment pipelines and incurring substantial computational overhead while neglecting the commonalities across heterogeneous tasks. In contrast, the proposed FreqDINO++ framework unifies heterogeneous ultrasound tasks within a single architecture, enabling universal analysis without maintaining separate task-specific networks.

B. Vision Foundation Models in Medical Imaging

The vision foundation models pre-trained on large-scale datasets have emerged as a promising paradigm for medical image analysis. The SAM series [28], [29] revolutionized segmentation through interactive prompting mechanisms, enabling remarkable generalization across diverse visual domains. Subsequent adaptations have tailored SAM to medical imaging, including MedSAM [30] for universal medical segmentation, Medical SAM Adapter [31] for parameter-efficient domain adaptation, and SAM2-Adapter [32] for extending SAM 2 to broader downstream tasks. In the ultrasound domain, USFM [2] constructed a multi-organ database containing over two million images and employed spatial-frequency dual masked image modeling to learn universal ultrasound representations for segmentation and classification.

Recent advances have also further explored prompt-driven and encoder-based adaptations specifically for ultrasound analysis. CC-SAM [33] incorporated cross-feature attention to handle low-contrast boundaries in ultrasound images, Lin *et al.* [34] designed an auto-prompting mechanism for end-to-end ultrasound segmentation, Zhang *et al.* [35] adapted vision foundation models for real-time ultrasound analysis, and Chen *et al.* [7] proposed a multi-organ foundation model with task prompts and anatomical priors. Meanwhile, DINOv3 [22] has demonstrated superior dense feature quality through self-supervised learning, inspiring adaptations such as DINO U-Net [25], SegDINO [26], and GuiDINO [36] that exploit its high-fidelity representations for medical segmentation. Despite the progress, these methods remain constrained by parameter-heavy adaptation. Different from these approaches, FreqDINO++ adapts DINOv3 through task-conditioned routing and frequency-guided enhancement, achieving parameter-efficient multi-task adaptation that coordinates dense and global prediction objectives within a unified framework.

III. Methods

A. Overview of FreqDINO++

We present FreqDINO++ as a unified framework for universal ultrasound image analysis in Fig. 2. Given an ultrasound image from the k-th task, we first feed it into a frozen DINOv3 ViT-Large backbone, where each Transformer block is equipped with a MR-Adapter to perform task-conditioned routing. The adapted patch tokens are then processed by the F^2-Enhancer, which progressively aggregates multi-scale frequency information to enhance feature representations. Finally, the enhanced features along with the CLS token are delivered to the TC-Decoder to jointly serve segmentation, classification, detection, and regression tasks.

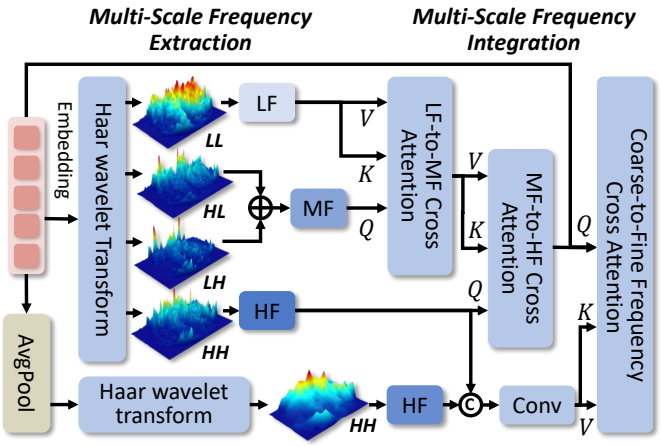

Fig. 3. The illustration of F²-Enhancer, decomposing adapted features into multiple frequency bands at two complementary scales via parallel Haar wavelet transforms and progressively integrating them through cascaded cross-band attention, providing frequency-enhanced representations for diverse downstream tasks.

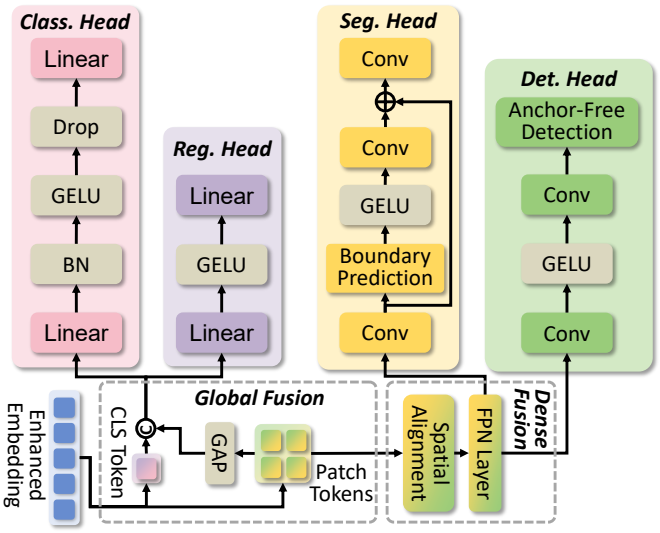

Fig. 4. The architecture of the TC-Decoder, which exploits the intrinsic alignment between task objectives and representation granularity: classification and regression are coupled through a global fusion pathway driven by semantic representations, while segmentation and detection are coupled through a dense fusion pathway driven by spatial representations, achieving harmonious multi-task collaboration.

In general, our FreqDINO++ is specifically designed to address the challenges of adapting vision foundation models to multi-task ultrasound analysis. The MR-Adapter enables efficient task-conditioned routing at every encoder layer without maintaining separate adapters for each task. The F²-Enhancer exploits the frequency-dependent characteristics of ultrasound images to enhance feature representations. Meanwhile, the TC-Decoder bridges dense and global tasks through token-level collaboration. Together, these modules enable FreqDINO++ to achieve universal ultrasound analysis within a single unified framework.

B. Multi-Task Routing Adapter

While DINOv3 [22] delivers powerful general-purpose representations, fully fine-tuning its heavy vision encoder for every ultrasound task is computationally prohibitive and quickly becomes intractable as the number of tasks grows. To address this, we devise the MR-Adapter that efficiently integrates task-common with task-specific knowledge within a single adapter module, enabling flexible multi-task adaptation in a parameter-efficient manner.

Specifically, the MR-Adapter is inserted after the feed-forward network in each Transformer block and employs a shared bottleneck architecture with deterministic task-conditioned routing. Given the input token sequence $\mathbf{x} \in \mathbb{R}^{B\times N\times C}$ and a task identifier $t \in \{\text{seg}, \text{cls}, \text{det}, \text{reg}\}$, the MR-Adapter first projects the input into a low-dimensional bottleneck space of dimension r through a shared down-projection $\mathbf{W}_{\text{down}} \in \mathbb{R}^{C\times r}$. Within this compact space, T lightweight task-specific transformation pathways $\{\mathbf{E}_t\}_{t=1}^{T}$ are maintained, each implemented as a MLP layer that captures task-specific adaptation patterns. The task identifier t deterministically selects the corresponding pathway $\mathbf{E}_t$ without any learnable gating or routing probability, ensuring stable training free from expert collapse. The transformed features are then mapped back to the original dimension via a shared up-projection $\mathbf{W}_{\text{up}} \in \mathbb{R}^{r\times C}$. This process is formulated as:

$$\mathbf{y} = \mathbf{x} + \mathbf{W}_{\text{up}} \cdot \sigma\Big(\mathbf{E}_t\big(\mathbf{W}_{\text{down}} \cdot \mathbf{x}\big)\Big), \tag{1}$$

where $\sigma(\cdot)$ denotes the GELU activation function applied after the task-specific routing. Since $\mathbf{W}_{\text{down}}$ and $\mathbf{W}_{\text{up}}$ are shared across all tasks, the MR-Adapter naturally aligns feature spaces of different tasks while the lightweight pathways $\{\mathbf{E}_t\}$ introduce only minimal task-specific parameters, achieving substantial parameter savings compared with maintaining independent per-task adapters. Through this design, the MR-Adapter achieves fine-grained task-aware adaptation from the earliest encoder layers, enabling effective cross-task knowledge sharing while preserving task-specific specialization for universal ultrasound analysis.

C. Frequency-Aware Feature Enhancer

Ultrasound images encode clinically significant information across distinct frequency bands: high-frequency components capture boundary details critical for segmentation and detection, mid-frequency components characterize tissue textures essential for classification, and low-frequency components describe global anatomical structures that underpin regression-based biometric measurements [9], [24]. However, the DINOv3 backbone, operating entirely in the spatial domain, cannot explicitly exploit such frequency-dependent characteristics. To this end, we propose the F²-Enhancer that fully exploits the multi-scale frequency characteristics inherent in ultrasound images to produce discriminative feature representations for diverse downstream tasks, as shown in Fig. 3.

Particularly, given the spatial feature map $\mathbf{F} \in \mathbb{R}^{B\times C\times H\times W}$ reshaped from the adapted patch tokens, we apply two parallel Haar wavelet transforms operating at the original and a coarser scale to extract multi-scale frequency information. The

Algorithm 1: Optimization Pipeline for FreqDINO++

Input: Ultrasound datasets $\{\mathcal{D}_t\}_{t=1}^{T}$; Frozen DINOv3 encoder E; MR-Adapter $\{\mathbf{W}_{\text{down}}, \mathbf{W}_{\text{up}}, \{\mathbf{E}_t\}_{t=1}^{T}\}$; F^2-Enhancer $\mathcal{E}$; TC-Decoder $\mathcal{D}_{\text{TC}}$

Output: Trained FreqDINO++ model

for *epoch* $= 1$ *to* N_{epoch} **do**
  Construct unified index pool $\mathcal{P} = \bigcup_{t=1}^{T} \mathcal{P}_t$;
  while $\mathcal{P}$ *is not exhausted* **do**
    Sample task $t \sim \text{Uniform}(\{1, \dots, T\})$;
    Sample a task-homogeneous mini-batch $\mathcal{B}_t$ from $\mathcal{P}_t$;
    for *each* $\mathbf{x} \in \mathcal{B}_t$ **do**
      $\mathbf{F}_{\text{patch}}, \mathbf{t}_{\text{cls}} = E(\mathbf{x};\, t)$ via Eq. (1);
      Reshape $\mathbf{F}_{\text{patch}}$ to $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$;
      $\mathbf{F}_{\text{enh}} = \mathcal{E}(\mathbf{F})$ via Eq. (2)-(5);
    end
    if $t \in \{seg, det\}$ **then**
      $\hat{y} = \mathcal{D}_{\text{TC}}^{\text{local}}(\mathbf{F}_{\text{enh}})$;
    else
      $\hat{y} = \mathcal{D}_{\text{TC}}^{\text{global}}(\mathbf{F}_{\text{enh}}, \mathbf{t}_{\text{cls}})$ via Eq. (6)-(7);
    end
    Compute $\mathcal{L}_{\text{FreqDINO}}$ and update parameters;
  end
end

original-scale branch directly decomposes $\mathbf{F}$ into four subbands $\{\mathbf{F}_{LL}, \mathbf{F}_{LH}, \mathbf{F}_{HL}, \mathbf{F}_{HH}\}$ and groups them into three frequency bands as follows:

$$\begin{aligned} \mathbf{F}_{\text{LF}} &= \phi_L(\mathbf{F}_{LL}), \\ \mathbf{F}_{\text{MF}} &= \phi_M(\mathbf{F}_{LH} + \mathbf{F}_{HL}), \\ \mathbf{F}_{\text{HF}} &= \phi_H(\mathbf{F}_{HH}), \end{aligned} \tag{2}$$

where ϕ_L, ϕ_M, ϕ_H are 1×1 convolutions for channel alignment, and the mid-frequency band is obtained by element-wise summation of the two directional detail subbands. In parallel, the coarse-scale branch first applies an average pooling operation $\mathbf{F}^{\downarrow} = \text{AvgPool}(\mathbf{F})$ and then performs Haar wavelet decomposition, retaining only the HH subband as $\mathbf{F}_{\text{HF}}^{c} = \phi_H^c(\mathbf{F}_{HH}^{\downarrow})$ to provide compact high-frequency structural cues over an enlarged receptive field, while discarding the coarse LL, LH, and HL subbands to avoid redundancy with the low- and mid-frequency context already preserved in the fine-scale branch. To enable effective cross-band interaction, we flatten each band into token sequences and design a cascaded cross-attention mechanism that progressively propagates contextual information from low to high frequencies. The mid-frequency tokens first attend to the low-frequency tokens to absorb global structural context, and the enriched mid-frequency tokens then serve as context for the high-frequency tokens:

$$\begin{aligned} \hat{\mathbf{Z}}_{\text{MF}} &= \mathbf{Z}_{\text{MF}} + \text{CA}(\mathbf{Z}_{\text{MF}}, \mathbf{Z}_{\text{LF}}), \\ \hat{\mathbf{Z}}_{\text{HF}} &= \mathbf{Z}_{\text{HF}} + \text{CA}(\mathbf{Z}_{\text{HF}}, \hat{\mathbf{Z}}_{\text{MF}}), \end{aligned} \tag{3}$$

where $\text{CA}(\mathbf{Q}, \mathbf{K})$ denotes the multi-head cross-attention with $\mathbf{Q}$ as query and $\mathbf{K}$ as both key and value. To further integrate the multi-scale boundary cues, we concatenate the enriched fine-scale and the coarse-scale high-frequency tokens and apply a 1×1 convolution to produce the fused multi-scale value, upon which a coarse-to-fine cross-attention is performed with the enriched fine-scale tokens as the query:

$$\begin{aligned} \mathbf{V}_{\text{cf}} &= \text{Conv}\big(\text{Cat}[\hat{\mathbf{Z}}_{\text{HF}},\, \mathbf{Z}_{\text{HF}}^{c}]\big), \\ \hat{\mathbf{Z}}_{\text{cf}} &= \hat{\mathbf{Z}}_{\text{HF}} + \text{CA}(\hat{\mathbf{Z}}_{\text{HF}},\, \mathbf{V}_{\text{cf}}). \end{aligned} \tag{4}$$

The integrated representation is then projected back to the spatial domain with a residual connection:

$$\mathbf{F}_{\text{enh}} = \mathbf{F} + \lambda \cdot \text{Proj}(\hat{\mathbf{Z}}_{\text{cf}}), \tag{5}$$

where λ is a learnable scaling factor. On this basis, our proposed F^2-Enhancer module explicitly captures multi-scale frequency information inherent in ultrasound images, enabling the enhanced representations to simultaneously encode boundary details, tissue textures, and anatomical structures that collectively benefit all downstream tasks.

D. Task-Aligned Collaborative Decoder

Existing ultrasound analysis methods typically employ independent networks to address different tasks [4], [5], [8], where dense prediction tasks (segmentation and detection) and global prediction tasks (classification and regression) are handled by entirely separate architectures without any feature interaction. This isolation not only introduces substantial computational redundancy but also prevents complementary knowledge transfer between related tasks, such as boundary-aware features benefiting both segmentation and detection, or global context supporting both classification and regression. To overcome this limitation, we design the TC-Decoder that leverages the collaborative interaction between local and global token representations to jointly serve all four tasks within a unified architecture, as shown in Fig. 4.

Specifically, the TC-Decoder receives the frequency-enhanced spatial features $\mathbf{F}_{\text{enh}}$ and the CLS token $\mathbf{t}_{\text{cls}} \in \mathbb{R}^{B\times C}$ from the encoder, and routes them into two complementary branches through a Dense Fusion module and a Global Fusion module, respectively. For the local branch serving dense prediction tasks, the Dense Fusion module first applies a Spatial Alignment operation that transforms the single-scale patch tokens into multi-scale feature pyramids with strides $\{4, 8, 16, 32\}$, followed by a Feature Pyramid Network (FPN) layer for hierarchical feature aggregation. The aggregated features are then delivered to a boundary-guided segmentation head that first predicts boundary maps and fuses them back into the main feature stream via residual summation before the final mask prediction, and an anchor-free detection head for center-point heatmap and bounding box regression. For the global branch serving holistic prediction tasks, the Global Fusion module fuses the spatial and semantic representations through CLS token-guided feature fusion:

$$\mathbf{h}_{\text{fused}} = \text{Cat}\big(\text{GAP}(\mathbf{F}_{\text{enh}}),\, \mathbf{t}_{\text{cls}}\big) \in \mathbb{R}^{B\times 2C}, \tag{6}$$

where $\text{GAP}(\cdot)$ denotes global average pooling that compresses the spatial features into a compact global representation, and $\text{Cat}(\cdot)$ concatenates it with the CLS token to form a

joint embedding that encodes both frequency-enhanced spatial information and high-level semantic knowledge. The fused representation is then fed into a classification head and a regression head as follows:

$$\begin{aligned} \hat{y}_{\text{cls}} &= \mathbf{W}_c \cdot \text{Dropout}\big(\text{BN}(\mathbf{h}_{\text{fused}})\big), \\ \hat{y}_{\text{reg}} &= \mathbf{W}_r \cdot \mathbf{h}_{\text{fused}}, \end{aligned} \tag{7}$$

where $\mathbf{W}_c$ and $\mathbf{W}_r$ are linear projection weights, with the classification branch further applying batch normalization (BN) and dropout for regularization, while the regression branch directly projects the fused features to preserve coordinate precision. In this way, our TC-Decoder enables dense and global tasks to share the same frequency-enhanced representations while maintaining task-appropriate decoding pathways, facilitating complementary knowledge transfer across heterogeneous tasks for universal ultrasound analysis.

E. Training and Optimization Pipeline

We summarize the training and optimization pipeline of FreqDINO++ for universal ultrasound analysis in Algorithm 1. To construct our framework, we initialize the frozen DINOv3 encoder E, the MR-Adapter parameters $\{\mathbf{W}_{\text{down}}, \mathbf{W}_{\text{up}}, \{\mathbf{E}_t\}_{t=1}^{T}\}$, the F^2-Enhancer $\mathcal{F}_{\text{F}^2\text{-Enhancer}}$, and the TC-Decoder $\mathcal{D}_{\text{TC}}$. The framework adopts a task-homogeneous batch training strategy, where each mini-batch contains data from only a single task to avoid gradient conflicts caused by heterogeneous loss scales, while a uniform task sampler ensures balanced training across all tasks within each epoch.

For each training iteration, a task identifier t is first uniformly sampled from the task pool, and a mini-batch $\mathcal{B}_t$ is drawn from the corresponding dataset $\mathcal{D}_t$. The input images are encoded through the frozen DINOv3 backbone with the MR-Adapter performing task-conditioned routing at each Transformer block. The resulting patch tokens are reshaped into spatial feature maps and enhanced by the F^2-Enhancer through cascaded wavelet decomposition and progressive cross-attention aggregation. The TC-Decoder then routes the enhanced features to the appropriate branch: the local branch for segmentation and detection, or the global branch for classification and regression.

The training is task-specific, with each task employing a tailored loss function. The overall objective of our FreqDINO++ framework can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{FreqDINO}} = \mathcal{L}_{\text{Dice}} + \lambda_{\text{seg}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{cls}}\mathcal{L}_{\text{CE}} \\ + \lambda_{\text{det}}\mathcal{L}_{\text{CenterNet}} + \lambda_{\text{reg}}\mathcal{L}_{\text{MSE}}, \end{aligned} \tag{8}$$

where $\mathcal{L}_{\text{BCE}}$ denotes the binary cross-entropy loss computed on morphologically generated boundary labels, and λ_{seg}, λ_{cls}, λ_{det}, λ_{reg} are weighting factors that balance the gradient magnitudes across tasks. Note that in each training iteration, only the loss terms corresponding to the sampled task t are activated, while the remaining terms are set to zero. By alternating optimization across the four tasks under this pipeline, our FreqDINO++ framework achieves comprehensive ultrasound image analysis across segmentation, classification, detection, and regression.

IV. Experiments

A. Datasets and Implementations

1) Datasets: To validate the effectiveness of the proposed FreqDINO++, we conduct comprehensive evaluations across three experimental settings: (1) multi-task evaluation on a large-scale challenge dataset, (2) single-task evaluation on individual benchmark datasets, and (3) generalization evaluation on an unseen dataset. Specifically, we adopt the FMC-UIA 2026 challenge dataset [40]–[43] as the primary multi-task benchmark, which provides tens of thousands of ultrasound images from multi-center cohorts covering 27 subtasks across four task categories. For single-task evaluation, we employ BUSI [44] for segmentation and classification, TN5000 [45] for detection, and FHC [46] for regression. Moreover, we evaluate the generalization capability of FreqDINO++ on the unseen TN3K [47] dataset. The details are as follows:

FMC-UIA [40]–[43] is a large-scale multi-task ultrasound image analysis challenge dataset from multi-center cohorts, covering 27 subtasks across four task categories: segmentation (12 subtasks, 15,991 images) targeting structures such as fetal head, fetal lung, breast tumor, and thyroid nodule; classification (9 subtasks, 16,362 images) including standard plane recognition, tumor classification, and organ identification; detection (3 subtasks, 4,333 images) for thyroid nodule, uterine fibroid, and spinal cord; and regression (3 subtasks, 3,702 images) for angle of progression, cervical length, and fetal femur length measurement.

BUSI [44] is a breast ultrasound image dataset containing 780 images from 600 female patients, categorized into three classes: normal (133 images), benign (437 images), and malignant (210 images). Each image is provided with pixel-level segmentation masks, making it suitable for both segmentation and classification tasks.

TN5000 [45] is a large-scale open-access thyroid nodule ultrasound dataset comprising 5,000 B-mode images with bounding box annotations and biopsy-confirmed labels.

FHC [46] is a fetal head circumference measurement dataset from the HC18 challenge, containing 1,334 two-dimensional ultrasound images annotated with ellipse fittings for head circumference estimation.

TN3K [47] is a thyroid nodule segmentation dataset containing 3,493 ultrasound images with pixel-level annotations. We use TN3K as an unseen dataset to evaluate the generalization capability of FreqDINO++ on the segmentation task.

2) Implementation Details: We perform all experiments on four NVIDIA RTX 5880 Ada GPUs with 48GB using the PyTorch framework. To enable comparison under a unified multi-task setting, we equipped competing backbones with task-specific output heads following their original design principles, and trained them under matched optimization settings. We utilize the pre-trained DINOv3 ViT-Large [22] as the frozen backbone encoder of our FreqDINO++. We apply the AdamW optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 0.05, and employ the cosine annealing strategy to gradually decay the learning rate to 1×10^{-6}. The loss weighting factors in Eq. (8) are set as λ_{seg}=0.3, λ_{cls}=1.0, λ_{det}=0.25, and λ_{reg}=50.0 to balance the gradient magnitudes

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON MULTI-TASK ULTRASOUND ANALYSIS ON THE FMC-UIA DATASET.

Methods	Segmentation			Classification			Detection			Regression		
	DSC↑	ASD↓	HD↓	AUC↑	F1↑	MCC↑	IoU↑	mAP↑	AP_{50}↑	MRE↓	MAE↓	SDR↑
AAU-Net [4]	70.91	16.75	54.53	82.78	60.23	48.10	50.92	18.84	48.62	17.40	34.66	32.28
TransUNet [23]	78.08	13.94	51.80	89.39	73.63	63.89	54.08	18.19	54.94	19.88	43.74	24.87
MADGNet [24]	75.14	16.12	56.56	89.94	65.88	55.21	57.59	24.30	65.32	18.48	31.92	32.46
SAM2 [29]	84.42	9.67	31.21	77.40	46.37	33.32	52.35	22.66	48.26	21.54	40.69	20.18
Med-SA [31]	88.43	7.37	21.75	90.22	78.63	70.13	56.90	20.78	63.98	26.61	59.12	16.08
SAM2-Adapter [32]	84.23	9.95	32.33	88.50	62.47	52.80	53.27	20.52	47.14	14.96	27.16	41.12
SAMUS [34]	80.67	13.21	49.17	90.18	75.43	67.66	62.71	31.90	64.09	22.40	40.13	29.51
UltraSAM [37]	87.42	7.89	24.48	89.77	68.95	58.73	55.76	20.46	55.95	18.37	37.6	24.99
Nora [38]	87.53	7.53	23.77	90.41	74.93	65.64	57.47	23.40	57.96	16.96	37.56	29.59
USFM [2]	86.96	7.62	24.25	87.85	73.30	63.18	46.78	22.93	48.13	12.33	23.93	54.57
LGFFM [9]	86.99	7.54	24.02	90.38	74.42	66.07	69.30	38.53	77.78	14.91	32.40	36.10
DINOv3-FD [39]	82.81	10.46	35.30	86.79	76.15	61.76	61.07	31.25	57.51	12.45	25.41	49.57
FreqDINO++	**88.46**	**6.62**	**20.32**	**91.24**	**81.57**	**71.99**	**74.83**	**47.75**	**82.84**	**7.74**	**14.21**	**71.69**

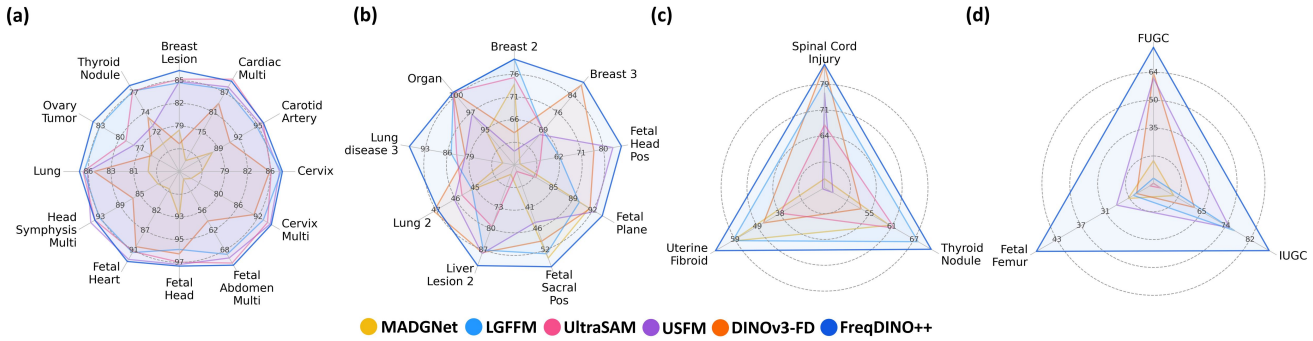

Fig. 5. Per-subtask comparison of FreqDINO++ against five representative baselines on the FMC-UIA dataset across all 27 clinical task scenarios, organized by task category: (a) segmentation (DSC, 12 subtasks), (b) classification (F1, 9 subtasks), (c) detection (mAP, 3 subtasks), and (d) regression (SDR, 3 subtasks). FreqDINO++ consistently encloses the polygons of competing methods across diverse anatomies and tasks.

across different tasks. We split all datasets into the training, validation, and test sets as 8:1:1, and all input images are resized to 256×256. The batch size and the training epoch are set to 32 and 150, respectively. For data augmentation, the segmentation task employs horizontal flip, vertical flip, random brightness contrast, and Gaussian noise, while the other three tasks apply random brightness contrast and Gaussian noise.

3) Evaluation Metrics: To perform a comprehensive evaluation of universal ultrasound analysis, we adopt the following standard metrics for each task category. For segmentation, we employ the Dice Similarity Coefficient (DSC), Average Surface Distance (ASD), and Hausdorff Distance (HD) to assess region overlap, average boundary deviation, and maximum boundary error, respectively. For classification, we utilize the Area Under the ROC Curve (AUC), F1 Score, and Matthews Correlation Coefficient (MCC) to evaluate discriminative ability, precision-recall balance, and overall prediction quality. For detection, we adopt Intersection over Union (IoU), mean Average Precision (mAP), and Average Precision at IoU threshold of 0.5 (AP_{50}) to quantify localization accuracy and detection performance. For regression, we employ the Mean Radial Error (MRE), Mean Absolute Error (MAE), and Successful Detection Rate (SDR) to measure positional deviation and the proportion of predictions within a clinically acceptable distance threshold. In the comparison results, the best performance values are highlighted in **bold**.

B. Comparison on Multi-Task Ultrasound Analysis

To comprehensively evaluate the performance of FreqDINO++, we first compare FreqDINO++ with state-of-the-art methods on the FMC-UIA multi-task benchmark. As shown in Table I, classical task-specific methods such as AAU-Net [4] and TransUNet [23] underperform foundation model-based approaches due to the absence of shared multi-task representations. Among foundation models, SAM-based methods (SAMUS [34], UltraSAM [37]) excel at segmentation but lag on classification, detection, and regression owing to their dense-prediction-oriented designs. Ultrasound-tailored foundation models perform stronger on individual tasks, with USFM [2] attaining the best baseline regression (12.33 MRE), LGFFM [9] the best detection (38.53% mAP), and DINOv3-FD [39] confirming the transferability of DINOv3 features. Yet all these methods treat each task in isolation, leaving cross-task synergies and frequency-domain cues unexploited.

In contrast, FreqDINO++ delivers the best results across all four task categories simultaneously. It achieves 88.46% DSC in segmentation with the lowest ASD (6.62) and HD (20.32), reducing them by 0.75 and 1.43 over Med-SA [31]. The gain is more pronounced in classification (91.24% AUC, 81.57% F1, 71.99% MCC), surpassing Nora [38] by 6.64% F1, and in detection (74.83% IoU, 47.75% mAP, 82.84% AP_{50}), outperforming LGFFM by 9.22% mAP. For regression,

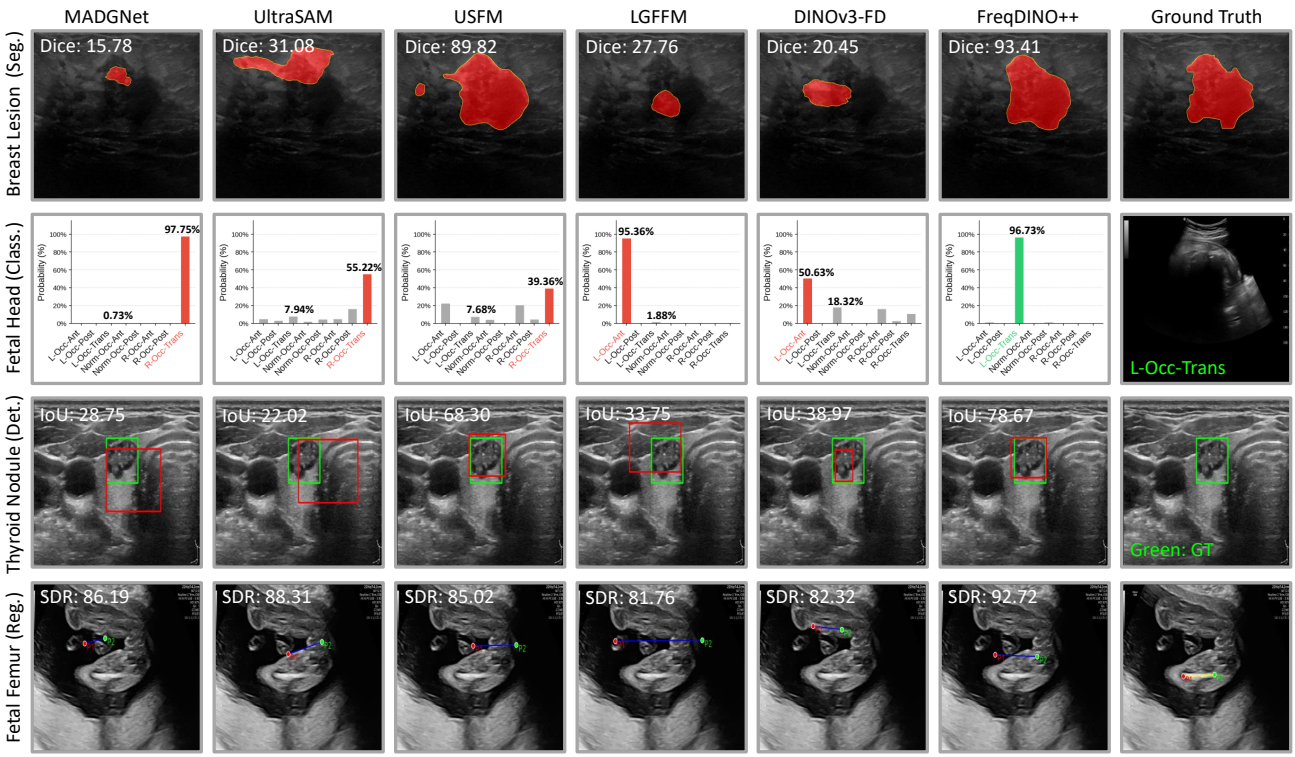

Fig. 6. Visualization of qualitative results across four tasks on the FMC-UIA dataset. Our FreqDINO++ exhibits the best results, producing more accurate lesion boundaries, correct diagnostic categories with higher confidence, tighter detection boxes, and more precise biometric landmarks.

it attains 7.74 MRE, 14.21 MAE, and 71.69% SDR, reducing MRE by 4.59 over USFM. The radar plot in Fig. 5 further shows that the polygon of FreqDINO++ consistently encloses those of leading competitors across all 27 subtasks, while the qualitative cases in Fig. 6 corroborate its superiority across heterogeneous tasks. These results confirm the effectiveness of our frequency-guided adaptation and multi-task routing for universal ultrasound analysis.

C. Comparison on Single-Task Benchmarks

We further evaluate FreqDINO++ on four external single-task benchmarks to assess its transferability beyond the multi-task training setting. As reported in Table II, classical task-specific methods exhibit limited performance, with AAU-Net [4] and MADGNet [24] obtaining only 71.31% and 67.32% DSC on BUSI segmentation. Foundation model-based approaches achieve stronger task-specific results: UltraSAM [37] reaches 79.97% DSC on BUSI segmentation, DINOv3-FD [39] attains 97.82% AUC on BUSI classification, LGFFM [9] obtains 42.89% mAP on TN5000 detection, and Nora [38] achieves 42.92% SDR on FHC regression. However, no single baseline performs consistently across all four tasks.

In contrast, our FreqDINO++ achieves state-of-the-art performance across all four single-task benchmarks simultaneously. On BUSI segmentation, FreqDINO++ attains 81.79% DSC and reduces HD to 31.86, surpassing UltraSAM by 1.82% in DSC. For BUSI classification, it obtains 97.91% AUC and 89.69% F1, outperforming Nora by 2.22% in F1. On TN5000 detection, FreqDINO++ reaches 46.98% mAP and 85.41% AP_{50}, improving over LGFFM by 4.09% in mAP. For FHC regression, it achieves the lowest MRE of 12.37 and the highest SDR of 47.19%, surpassing Nora by 4.27%. Qualitative comparisons of representative cases across the four task categories are further illustrated in Fig. 7. These results confirm that, despite being trained within a unified multi-task framework, the proposed FreqDINO++ framework delivers consistently superior performance across diverse external benchmarks and outperforms task-specific methods.

D. Ablation Study

To investigate the effectiveness of the MR-Adapter, F^2-Enhancer, and TC-Decoder, we conduct a comprehensive ablation study on the FMC-UIA multi-task dataset, as illustrated in Table III. The frozen DINOv3 backbone with vanilla linear decoders (1^{st} row) serves as the ablation baseline. By separately introducing the MR-Adapter (2^{nd} row), F^2-Enhancer (3^{rd} row), and TC-Decoder (4^{th} row), segmentation DSC is consistently improved over the baseline by 0.76%, 1.38%, and 0.64%, respectively. Among them, the F^2-Enhancer yields the largest segmentation gain (1.38% DSC) and regression refinement (1.18 MRE reduction), the TC-Decoder contributes the most to detection (1.34% mAP), and the MR-Adapter brings the largest regression SDR improvement (4.05%).

We further examine pairwise combinations. By comparing the 6^{th} and 7^{th} rows with the corresponding single-module configurations, the TC-Decoder consistently amplifies the benefits of both the MR-Adapter and F^2-Enhancer, particularly boosting detection mAP by 3.90% and regression SDR by 6.87%. Building upon these complementary effects, our complete FreqDINO++ (8^{th} row) integrates all three modules to

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS ON SINGLE-TASK ULTRASOUND BENCHMARKS.

Methods	BUSI-Seg			BUSI-Class			TN5000			FHC		
	DSC↑	ASD↓	HD↓	AUC↑	F1↑	MCC↑	IoU↑	mAP↑	AP_{50}↑	MRE↓	MAE↓	SDR↑
AAU-Net [4]	71.31	18.94	59.51	85.63	69.93	53.12	67.71	41.74	78.63	21.19	40.79	29.58
TransUNet [23]	72.51	21.77	73.75	95.23	84.00	75.66	68.34	40.30	79.12	15.44	34.77	37.08
MADGNet [24]	67.32	24.17	84.07	94.79	85.54	77.16	68.50	39.44	79.32	19.38	39.86	31.67
SAM2 [29]	75.93	20.52	79.12	80.98	62.62	42.19	50.03	15.19	44.19	26.38	46.32	22.19
Med-SA [31]	78.01	20.09	67.06	97.74	85.39	77.38	62.37	27.11	63.97	13.75	32.07	38.85
SAM2-Adapter [32]	77.81	17.39	57.82	95.30	81.91	71.00	54.43	19.71	52.12	16.34	34.65	36.67
SAMUS [34]	60.46	34.03	111.81	91.90	85.47	74.40	68.66	35.35	78.39	14.31	31.73	41.15
UltraSAM [37]	79.97	12.78	42.98	96.16	84.35	75.61	61.25	24.42	59.24	15.20	31.67	40.42
Nora [38]	76.76	15.68	61.46	97.25	87.47	77.63	60.71	25.08	62.83	14.79	30.16	42.92
USFM [2]	79.35	16.50	54.30	78.15	55.32	30.46	64.52	34.26	74.46	22.20	45.45	25.52
LGFFM [9]	78.72	15.42	46.82	96.64	84.90	72.64	72.43	42.89	76.01	14.63	32.20	38.12
DINOv3-FD [39]	71.98	17.20	64.18	97.82	86.66	79.15	57.23	23.63	62.24	17.23	36.52	35.73
FreqDINO++	**81.79**	**9.84**	**31.86**	**97.91**	**89.69**	**83.41**	**72.69**	**46.98**	**85.41**	**12.37**	**28.26**	**47.19**

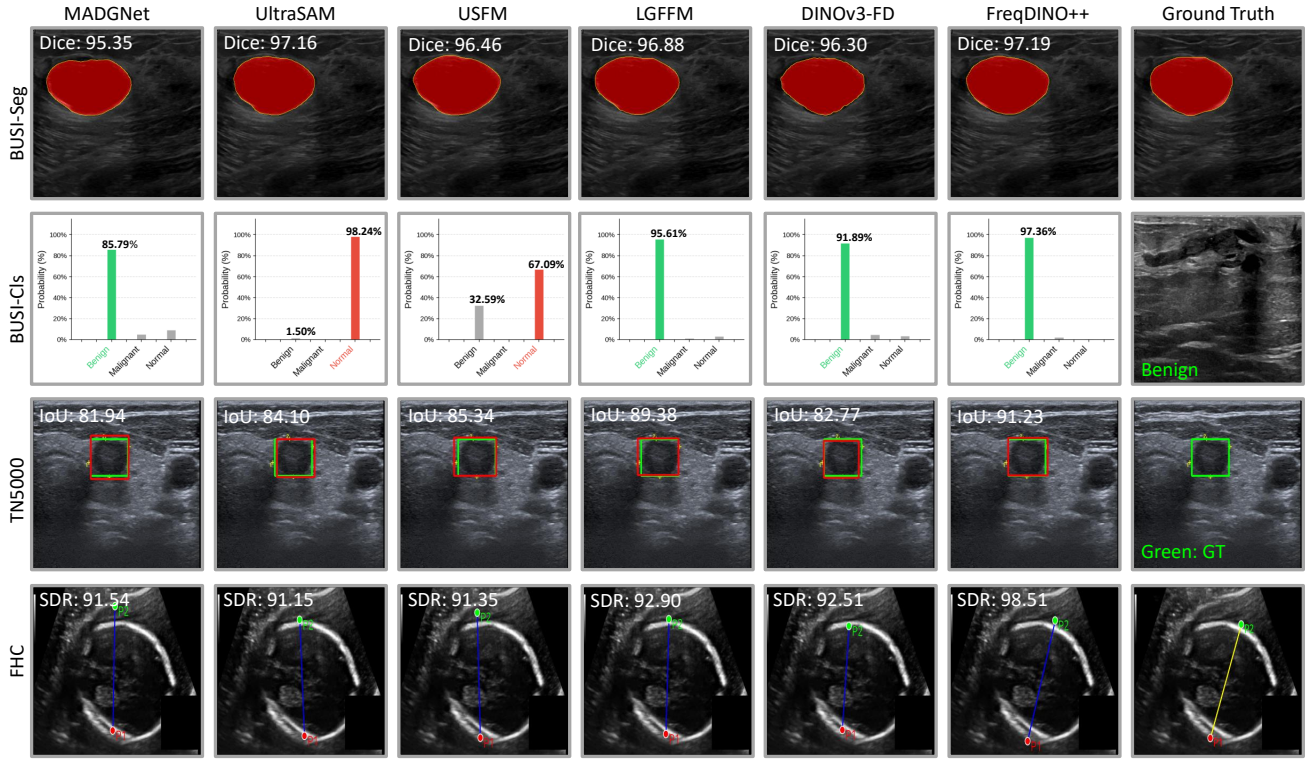

Fig. 7. Visualization of qualitative results on four benchmarks. Our FreqDINO++ consistently achieves the best results across all four single-task settings, with more accurate masks, correct categories, tighter detection boxes, and more precise biometric landmarks than competing methods.

achieve the best performance across all tasks, with overall gains of 2.97% DSC, 12.03% F1, 5.93% mAP, and 12.44% SDR over the baseline. These comprehensive ablation experiments validate that the tailored MR-Adapter, F^2-Enhancer, and TC-Decoder collectively contribute to the superior performance of FreqDINO++ by enabling parameter-efficient task-conditioned adaptation, capturing multi-scale frequency characteristics, and coordinating dense and global prediction tasks through token-level collaboration.

E. Effectiveness of Frequency-aware Feature Enhancer

To evaluate the contribution of each frequency band in the F^2-Enhancer, we progressively incorporate different frequency components through cascaded wavelet decomposition, as shown in Table IV. Without any frequency decomposition, the framework operates entirely in the spatial domain and achieves baseline performance across all four tasks (1^{st} row). Introducing the low-frequency band alone (2^{nd} row) yields consistent gains across all tasks, with 0.30% AUC improvement and 1.91% MCC gain in classification, as the LF band captures global anatomical structures that benefit holistic prediction. Sequentially adding the mid-frequency band (3^{rd} row) further boosts classification with a 3.66% F1 gain, confirming that tissue texture patterns encoded in the MF band are critical for diagnostic categorization. The complete three-band integration (4^{th} row), incorporating the high-frequency

TABLE III
ABLATION STUDY OF FREQDINO++ ON THE FMC-UIA MULTI-TASK DATASET.

MR-Adapter	F^2-Enhancer	TC-Decoder	Segmentation			Classification			Detection			Regression		
			DSC↑	ASD↓	HD↓	AUC↑	F1↑	MCC↑	IoU↑	mAP↑	AP_{50}↑	MRE↓	MAE↓	SDR↑
			85.49	8.39	26.33	89.94	69.54	60.75	70.11	41.82	77.33	10.24	19.47	59.25
✓			86.25	7.74	23.46	89.57	73.07	62.76	71.12	40.26	82.14	9.52	17.38	63.30
	✓		86.87	7.64	24.43	90.14	74.42	63.25	72.05	40.83	81.53	9.06	17.25	62.11
		✓	86.13	7.92	23.58	90.01	73.28	62.61	71.81	43.16	82.22	9.71	18.20	61.79
✓	✓		87.16	7.31	22.03	90.71	76.20	65.40	72.53	43.47	83.03	8.51	15.76	64.86
✓		✓	87.06	7.52	22.26	90.80	78.30	66.80	73.25	44.16	82.63	8.61	15.22	69.56
	✓	✓	87.79	6.97	21.73	90.63	78.64	68.15	72.77	43.47	79.99	8.48	15.76	68.98
✓	✓	✓	**88.46**	**6.62**	**20.32**	**91.24**	**81.57**	**71.99**	**74.83**	**47.75**	**82.84**	**7.74**	**14.21**	**71.69**

TABLE IV
EFFECTIVENESS OF FREQUENCY-AWARE FEATURE ENHANCER WITH DIFFERENT FREQUENCY BAND COMBINATIONS ON THE FMC-UIA DATASET.

LF	MF	HF	Segmentation			Classification			Detection			Regression		
			DSC↑	ASD↓	HD↓	AUC↑	F1↑	MCC↑	IoU↑	mAP↑	AP_{50}↑	MRE↓	MAE↓	SDR↑
			87.15	7.44	21.98	90.21	77.17	65.97	73.07	44.62	81.49	8.44	15.47	67.59
✓			87.47	7.22	21.86	90.51	77.34	67.88	73.58	46.07	82.70	8.39	15.01	68.24
✓	✓		88.29	6.77	20.48	91.03	81.00	70.70	74.40	46.06	82.52	8.23	15.23	68.95
✓	✓	✓	**88.46**	**6.62**	**20.32**	**91.24**	**81.57**	**71.99**	**74.83**	**47.75**	**82.84**	**7.74**	**14.21**	**71.69**

TABLE V
COMPARISON ON GENERALIZATION CAPABILITY ON THE UNSEEN TN3K SEGMENTATION DATASET.

Methods	DSC↑	ASD↓	HD↓
AAU-Net [4]	68.49	20.82	66.89
TransUNet [23]	76.68	15.48	50.58
MADGNet [24]	71.69	18.38	61.41
SAM2 [29]	76.25	15.25	47.82
Med-SA [31]	80.18	13.43	40.18
SAM2-Adapter [32]	75.75	16.12	54.06
SAMUS [34]	79.84	13.31	45.61
UltraSAM [37]	79.93	12.62	39.43
Nora [38]	80.81	12.13	39.14
USFM [2]	77.70	12.55	41.63
LGFFM [9]	80.34	12.76	42.83
DINOv3-FD [39]	74.42	16.94	62.18
FreqDINO++	**82.86**	**10.71**	**37.62**

band for boundary details, achieves optimal performance with 88.46% DSC and 6.62 ASD in segmentation, representing total gains of 1.31% DSC, 4.40% F1, 3.13% mAP, and 4.10% SDR over the spatial-only baseline. The progressive improvements across band combinations confirm that the LF, MF, and HF bands play distinct yet synergistic roles, jointly aligning with the structural, textural, and boundary-related demands of heterogeneous ultrasound tasks.

F. Comparison on Generalization Capability

To further validate the generalization capability of FreqDINO++, we evaluate all methods on the unseen TN3K thyroid nodule segmentation dataset, where all models directly apply the weights trained on the FMC-UIA multi-task dataset without any fine-tuning. As shown in Table V, classical methods such as AAU-Net [4] and MADGNet [24] suffer substantial performance degradation on this unseen dataset, achieving only 68.49% and 71.69% DSC, respectively, as their task-specific architectures lack the transferable representations needed for cross-dataset generalization. Foundation model-based approaches demonstrate comparatively stronger robustness, with Nora [38] and LGFFM [9] attaining 80.81% and 80.34% DSC, benefiting from their large-scale pre-trained features. In contrast, our FreqDINO++ achieves the best generalization performance with 82.86% DSC, 10.71 ASD, and 37.62 HD, surpassing the best baseline Nora by 2.05% in DSC and reducing HD by 1.52. These results demonstrate that the frequency-guided adaptation and multi-task routing mechanisms in FreqDINO++ not only enhance in-domain performance but also foster more transferable representations on unseen ultrasound datasets, confirming its significant potential for deployment in real-world clinical scenarios.

V. CONCLUSION

In this work, we propose FreqDINO++, a frequency-guided multi-task routing framework for universal ultrasound analysis to address the challenges of adapting vision foundation models to heterogeneous ultrasound tasks. Built on a frozen DINOv3 backbone, FreqDINO++ integrates a MR-Adapter to efficiently balance task-common and task-specific knowledge through deterministic task-conditioned routing, and a F^2-Enhancer that decomposes features into multiple frequency bands via cascaded wavelet transforms and progressively aggregates them to capture rich frequency information inherent in ultrasound images. Additionally, a TC-Decoder leverages the collaborative interaction between local and global token representations to jointly serve segmentation, classification, detection, and regression tasks within a unified dual-branch architecture. Extensive experiments on diverse datasets show that FreqDINO++ consistently surpasses state-of-the-art methods across all four task categories, i.e., segmentation, classification, detection, and regression settings, while also exhibiting strong generalization capability on unseen datasets.

REFERENCES

[1] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Med. Image Anal.*, vol. 79, p. 102444, 2022.

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

[3] Z. Lin, H. Zhang, X. Duan, Y. Bai, J. Wang, Q. Liang, J. Zhou, F. Xie, Z. Shentu, R. Huang *et al.*, "Deep learning approach for screening neonatal cerebral lesions on ultrasound in china," *Nat. Commun.*, vol. 16, no. 1, p. 7778, 2025.

[4] G. Chen, L. Li, Y. Dai, J. Zhang, and M. H. Yap, "Aau-net: an adaptive attention u-net for breast lesions segmentation in ultrasound images," *IEEE Trans. Med. Imaging*, vol. 42, no. 5, pp. 1289–1300, 2022.

[5] X. Tao, Y. Cao, Y. Jiang, X. Wu, D. Yan, W. Xue, S. Zhuang, X. Yang, R. Huang, J. Zhang *et al.*, "Enhancing lesion detection in automated breast ultrasound using unsupervised multi-view contrastive learning with 3d detr," *Med. Image Anal.*, vol. 101, p. 103466, 2025.

[6] C. Qin, J. Cao, F. S. Khan, S. Khan, H. Fu, E. Ahissar, and R. M. Anwer, "Real-time breast lesion detection in videos via spatial-temporal feature aggregation," in *Medical Imaging with Deep Learning*, 2025.

[7] H. Chen, Y. Cai, C. Wang, L. Chen, B. Zhang, H. Han, Y. Guo, H. Ding, and Q. Zhang, "Multi-organ foundation model for universal ultrasound image segmentation with task prompt and anatomical prior," *IEEE Trans. Med. Imaging*, vol. 44, no. 2, pp. 1005–1018, 2024.

[8] K. W. Mikołaj, A. N. Christensen, C. A. Taksøe-Vester, A. Feragen, O. B. Petersen, M. Lin, M. Nielsen, M. B. S. Svendsen, and M. G. Tolsgaard, "Predicting abnormal fetal growth using deep learning," *npj Digit. Med.*, vol. 8, no. 1, p. 318, 2025.

[9] X. Luo, Y. Wang, and L. Ou-Yang, "Lgffm: A localized and globalized frequency fusion model for ultrasound image segmentation," *IEEE Trans. Med. Imaging*, 2025.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[11] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[12] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[13] W. Zhang, Y. Cao, X. Hu, G. Sun, Y. Zhang, Y. Li, Q. Cai, and Z. Liu, "Annular prior prompt learning for medical images segmentation," *IEEE Trans. Biomed. Eng.*, 2025.

[14] D. Wang, T. Zhou, S. Gao, and J. Yang, "Echo flow-induced temporal correlation learning for ultrasound video object segmentation," *IEEE Trans. Biomed. Eng.*, 2025.

[15] Y. Wang, W. Lu, L. Xu, H. Xu, and D. Kong, "Deep learning-based ultrasound diagnostic model for follicular thyroid carcinoma," *Eur. Radiol.*, vol. 36, no. 1, pp. 357–366, 2026.

[16] T. Jiang, Y. Li, Y. Li, W. Xing, M. Yu, F. Xie, and D. Ta, "A segmentation knowledge-based global-local attention network for tumor classification in breast ultrasound images," *Pattern Recognit.*, p. 112152, 2025.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Springer, 2020, pp. 213–229.

[19] C. Zhu, J. Li, M. Ren, Y. Chen, Y. Chen, M. Li, Q. Cai, T. Wang, Z. Wang, H. Song *et al.*, "Machine learning-enhanced prediction of fetal growth restriction using fetal cardiac remodeling parameters," *BMC Med.*, vol. 23, no. 1, p. 634, 2025.

[20] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197–207, 2019.

[21] N. Ibtehaz and D. Kihara, "Acc-unet: A completely convolutional unet model for the 2020s," in *MICCAI*. Springer, 2023, pp. 692–702.

[22] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[23] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang *et al.*, "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Med. Image Anal.*, vol. 97, p. 103280, 2024.

[24] J.-H. Nam, N. S. Syazwany, S. J. Kim, and S.-C. Lee, "Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention," in *CVPR*, 2024, pp. 11 480–11 491.

[25] Y. Gao, H. Li, F. Yuan, X. Wang, and X. Gao, "Dino u-net: Exploiting high-fidelity dense features from foundation models for medical image segmentation," *arXiv preprint arXiv:2508.20909*, 2025.

[26] S. Yang, H. Wang, Z. Xing, S. Chen, and L. Zhu, "Segdino: An efficient design for medical and natural image segmentation with dino-v3," *arXiv preprint arXiv:2509.00833*, 2025.

[27] Y. Zhang, Q. Xu, Y. Li, X. He, Q. Zhang, M. Haque, R. Qu, W. Duan, and Z. Chen, "Freqdino: Frequency-guided adaptation for generalized boundary-aware ultrasound image segmentation," *ISBI*, 2026.

[28] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *ICCV*, 2023, pp. 4015–4026.

[29] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," in *ICLR*, 2025.

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

[31] J. Wu, Z. Wang, M. Hong, W. Ji, H. Fu, Y. Xu, M. Xu, and Y. Jin, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *Med. Image Anal.*, vol. 102, p. 103547, 2025.

[32] T. Chen, A. Lu, L. Zhu, C. Ding, C. Yu, D. Ji, Z. Li, L. Sun, P. Mao, and Y. Zang, "Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more," in *ICLR 2025 Workshop*, 2025.

[33] S. N. Gowda and D. A. Clifton, "Cc-sam: Sam with cross-feature attention and context for ultrasound image segmentation," in *ECCV*. Springer, 2024, pp. 108–124.

[34] X. Lin, Y. Xiang, L. Yu, and Z. Yan, "Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting," in *MICCAI*. Springer, 2024, pp. 24–34.

[35] X. Zhang, E. Z. Chen, L. Zhao, X. Chen, Y. Liu, B. Maihe, J. S. Duncan, T. Chen, and S. Sun, "Adapting vision foundation models for real-time ultrasound image segmentation," in *MICCAI*. Springer, 2025, pp. 24–34.

[36] Z. Liang, W. Guo, J. Gan, Y. Song, R. Chen, H. Chang, and W. Cai, "Guidino: Rethinking vision foundation model in medical image segmentation," *arXiv preprint arXiv:2603.01115*, 2026.

[37] A. Meyer, A. Murali, F. Zarin, D. Mutter, and N. Padoy, "Ultrasam: a foundation model for ultrasound using large open-access segmentation datasets," *Int. J. Comput. Assist. Radiol. Surg.*, pp. 1–10, 2025.

[38] Z. Wei, C. Wu, H. Du, R. Yu, B. Du, and Y. Xu, "Noise-robust tuning of sam for domain generalized ultrasound image segmentation," in *MICCAI*. Springer, 2025, pp. 476–486.

[39] Z. He, Y. Fu, and Y. Jin, "Incentivizing dinov3 adaptation for medical vision tasks via feature disentanglement," in *MIDL*, 2026.

[40] Y. Lu, M. Zhou, D. Zhi, M. Zhou, X. Jiang, R. Qiu, Z. Ou, H. Wang, D. Qiu, M. Zhong *et al.*, "The jnu-ifm dataset for segmenting pubic symphysis-fetal head," *Data Brief*, vol. 41, p. 107904, 2022.

[41] Y. Chen, B. Deng, Z. Peng, Y. Zhou, P. Sun, J. Wan, and J. Bai, "Comt: Co-training mean teachers semi-supervised training framework for cervical segmentation," in *ISBI*. IEEE, 2025, pp. 1–5.

[42] Y. Zhang, J. Zhu, S. Long, J. Bai, Y. Lu, and G. Chen, "An automatic measurement of cervix dilation in intrapartum ultrasound image," *Signal Image Video P.*, vol. 19, no. 3, p. 229, 2025.

[43] G. Chen, J. Bai, Z. Ou, Y. Lu, and H. Wang, "Psfhs: intrapartum ultrasound image dataset for ai-based segmentation of pubic symphysis and fetal head," *Sci. Data.*, vol. 11, no. 1, p. 436, 2024.

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

[45] H. Zhang, Q. Liu, X. Han, L. Niu, and W. Sun, "Tn5000: An ultrasound image dataset for thyroid nodule detection and classification," *Sci. Data.*, vol. 12, no. 1, p. 1437, 2025.

[46] T. L. van den Heuvel, D. de Bruijn, C. L. de Korte, and B. v. Ginneken, "Automated measurement of fetal head circumference using 2d ultrasound images," *PloS one*, vol. 13, no. 8, p. e0200412, 2018.

[47] H. Gong, J. Chen, G. Chen, H. Li, G. Li, and F. Chen, "Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules," *Comput. Biol. Med.*, vol. 155, p. 106389, 2023.