Tuning Medical Foundation Models for Inner Ear Temporal CT Analysis with Plug-and-play Domain Knowledge Aggregator

Weixun Wan[1], **Xinyang Jiang**[2*], **Zilong Wang**[2], **Bei Li**[3], **Cairong Zhao**[1*]

[1]School of Computer Science and Technology, Tongji University
[2]Microsoft Research Asia, Microsoft
[3]Department of Otolaryngology-Head and Neck Surgery, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine

Abstract

High-resolution computed tomography (CT) is essential for diagnosing hearing loss and planning interventions such as cochlear implantation, as it provides detailed visualization of inner-ear anatomy. This paper focuses on advancing AI-based analysis of inner-ear CT scans to support clinical decision-making. However, a major challenge lies in the scarcity of annotated data, which limits the applicability of conventional supervised learning techniques. To address this, we present the first publicly available Children's Inner Ear CT Dataset (CIED), comprising 722 CT scans labeled for structural anomaly detection, postoperative hearing outcome prediction, and anatomical segmentation. In addition, we explore the use of medical foundation models to improve generalization in data-scarce scenarios. Existing parameter-efficient adaptation methods often fall short in two ways: they lack a unified mechanism to adapt across diverse foundation model architectures, and they are not specifically designed to incorporate domain expert knowledge of inner-ear anatomy and pathology. To overcome these limitations, we propose Domain Knowledge Guided Tuning (DKGT), a plug-and-play framework that introduces a unified adapter—Domain Knowledge Aggregator(DKA)—to inject radiomics-based anatomical features into foundation models via cross-attention. DKA supports various backbone types and preserves pretrained representations of foundation model while enabling multi-layer integration of expert knowledge. Extensive experiments across multiple tasks demonstrate that DKGT consistently outperforms state-of-the-art classification methods, achieving superior performance and generalizability on inner-ear CT analysis.

Code — https://github.com/Vill-Lab/DKA-innerear

1 Introduction

Hearing loss affects over 1.5 billion people worldwide, impairing communication, cognition, and quality of life. Its causes range from congenital anomalies to infections and age-related degeneration, requiring accurate diagnosis and personalized treatment. High-resolution temporal bone CT is widely used to visualize key inner ear structures, aiding in diagnosing malformations, otosclerosis, and preoperative planning for cochlear implantation(Harnsberger et al. 1987; Miyasaka et al. 2010). Recently, artificial intelligence (AI) has shown promise in automating CT analysis, improving segmentation, anomaly detection, and diagnosis.

Despite its clinical significance, AI-based analysis of inner ear medical imaging remains a relatively underexplored field, largely due to two major challenges. First, acquiring high-quality inner ear CT data for diagnostic or prognostic purposes is both costly and time-consuming. It demands precise anatomical annotations, expert radiological assessments, and often longitudinal follow-up data, making the creation of large-scale labeled datasets difficult. Consequently, the scarcity of publicly available datasets poses a significant barrier to the development of supervised deep learning models. As a result, most inner ear AI research focuses on dense prediction tasks such as segmentation and localization(Girum, Crehange, and Lalande 2021; Vaidyanathan et al. 2021; Zhang et al. 2020) rather than classification. Second, clinical decision-making for inner ear conditions heavily relies on specialized domain knowledge. In routine diagnostic workflows, physicians carefully examine specific anatomical landmarks—such as the cochlea, semicircular canals, or internal auditory canal—drawing on their expertise in otology and radiology to identify subtle structural abnormalities(Casselman et al. 2001; Yiin, Tang, and Tan 2011). Unlike clinicians, deep learning models typically require large volumes of annotated data to achieve similar levels of diagnostic accuracy. Yet, current AI approaches often fail to incorporate this expert-driven knowledge, resulting in models that lack interpretability, generalizability, and alignment with clinical reasoning.

To address the challenge of data scarcity in inner ear imaging, we introduce the Children's Inner Ear CT Dataset (CIED), the first publicly available dataset specifically curated to support a range of clinical and research tasks on inner-ear analysis, including structural anomaly diagnosis, prognosis prediction following cochlear implantation, and anatomical annotation. CIED consists of 722 CT scans focused on the inner ear. Among them, 108 scans are accompanied by postoperative outcome measures and expert annotations of key anatomical structures, supporting both hearing recovery prediction and auxiliary segmentation tasks. The remaining 614 scans are labeled across six categories

*Corresponding authors. Email: xinyangjiang@microsoft.com; zhaocairong@tongji.edu.cn

of structural anomalies, enabling evaluation of anomaly detection models. This comprehensive dataset provides a standardized foundation for benchmarking and advancing AI-driven methods in inner ear CT analysis.

In order to train robust AI models with limited training data, we leverage recent advancements in medical imaging foundation models such as CT-FM (Pai et al. 2025) and Merlin (Blankemeier et al. 2024), which have demonstrated strong generalizability across a wide range of medical imaging tasks. To the best of our knowledge, this work represents the first attempt to adapt foundation models specifically for inner ear CT analysis. However, this adaptation poses two key challenges. First, existing medical foundation models adopt diverse architectural designs, including transformer based frameworks, convolutional neural networks (CNNs), and U-Net variants, making it difficult to establish a unified and consistent fine-tuning strategy. Second, although recent parameter-efficient fine-tuning (PEFT) methods, such as adapters (Rebuffi, Bilen, and Vedaldi 2017) and LoRA (Hu et al. 2021), help mitigate catastrophic forgetting, they are not specifically tailored for inner-ear CT analysis and lack mechanisms to incorporate domain-specific knowledge of inner-ear anatomy and pathology, where expert-driven structural understanding is crucial for accurate diagnosis.

Building on the proposed datasets, we benchmark a range of state-of-the-art(SOTA) AI methods for inner ear CT analysis, including 3D classification models and medical imaging foundation models. In addition, we introduce a novel plug-and-play framework, Domain Knowledge Guided Tuning (DKGT), specifically designed for inner ear imaging tasks. DKGT introduces Domain Knowledge Aggregator (DKA) as a unified adapter compatible with diverse foundation model architectures, including CNNs, U-Nets, and Vision Transformers. To effectively incorporate clinical domain knowledge, DKA models radiologically relevant features of key anatomical structures and encodes them as domain knowledge tokens, which are injected into the foundation model to guide learning. To facilitate interaction between these domain knowledge tokens and the image tokens extracted from the backbone model, DKA allows image tokens to query the most relevant knowledge tokens and refines them via a residual connection. This enables hierarchical integration of inner ear–specific knowledge across multiple layers of the model, without altering the size of the feature maps, thus enhancing adaptability to various network architectures. Experimental results show that DKGT significantly enhances generalization on small-scale datasets and consistently outperforms existing SOTA classification approaches, PEFT techniques, and even full fine-tuning methods across multiple evaluation tasks.

2 Related Work

2.1 AI in Inner Ear Image Domain

Artificial intelligence (AI) has promising applications in otology, including hearing prediction, vestibular disorder management, and automated image analysis(You et al. 2020). However, AI research focusing on inner ear imaging remains limited. Early studies used classical machine learning, such as SVM-based analysis of cochlear nerve features for predicting cochlear implant outcomes(Lu et al. 2022), but relied on manual feature extraction. Deep learning methods have primarily focused on segmentation: LFB-Net combined feed-forward and FCN feedback systems for micro-CT images(Girum, Crehange, and Lalande 2021); 3D U-Net models achieved expert-level MRI segmentation(Vaidyanathan et al. 2021; Fauser et al. 2019; Neves et al. 2021); and HeadLocNet localized inner-ear landmarks in CT scans(Zhang et al. 2020). These studies show feasibility but are limited by small datasets and task-specific designs; we provide a larger dataset and evaluate advanced deep learning and foundation models for inner ear analysis.

2.2 Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) methods allow large pretrained models to adapt to new tasks while using fewer computational resources. Adapter methods add small trainable modules between transformer layers, keeping the original model weights frozen. These approaches work well in both Natural Language Processing(Stickland and Murray 2019; Pfeiffer et al. 2020) and Computer Vision(Rebuffi, Bilen, and Vedaldi 2017; Chen et al. 2022; Fan et al. 2024; Song et al. 2023; Cen et al. 2025). Low-Rank Adaptation (LoRA) uses matrix decomposition to reduce the number of trainable parameters needed for fine-tuning(Hu et al. 2021; Mao et al. 2024; Zhu et al. 2024).

Another approach is prompt-based tuning, which modifies the input to the model rather than the model parameters themselves. First developed for NLP tasks(Li and Liang 2021; Lester, Al-Rfou, and Constant 2021; Liu et al. 2021), prompt tuning has been applied to vision and multimodal models(Zhou et al. 2022; Ge et al. 2023; Ju et al. 2022; Yao et al. 2024; Jia et al. 2022; Khattak et al. 2023). While these methods achieve good performance with minimal parameter updates, they do not incorporate medical domain knowledge, which motivates our approach for inner ear imaging.

3 Children's Inner Ear CT Dataset

We present the Children's Inner Ear CT Dataset (CIED), the first publicly available benchmark specifically for inner ear CT imaging. CIED supports three tasks: postoperative hearing recovery prediction, structural anomaly detection, and an auxiliary inner ear anatomical segmentation task. The dataset primarily consists of infants and young children, including 443 cases under 1 year, 229 aged 1–3 years, 40 aged 3–6 years, and 10 above 6 years, collected from 2017 to 2024. All CT volumes were rigidly registered with SimpleITK and cropped around the inner-ear ROI to a uniform size of 128×128×64.

Postoperative Hearing Recovery Prediction This task uses CT scans of cochlear implant patients along with hearing thresholds measured 12 months post-surgery. Average thresholds at 0.5, 1, 2, and 4 kHz are used to binarize outcomes into good or poor recovery. The task provides paired imaging and outcome labels for model training and evaluation. The statistical information of the hearing recovery prediction task is shown in Table 1.

Task	Gender(M/F)	Age(Months)	Slice Thickness	Number of Slices	Label Description	Label 0/1
Recovery Pred	71/37	19.22±19.37	0.6mm	6912	positive / negative outcome (0/1)	49/59
Structural Anomaly Classification	368/246	10.68±7.32	0.6mm	39296	overally normal/abnormal (0/1) not NBCNC/ NBCNC (0/1) not NIAC/ NIAC (0/1) not Mondini/ Mondini (0/1) not EVA/ EVA (0/1) not VSCD/ VSCD (0/1) not Michel/ Michel (0/1)	392/222 496/118 578/36 559/55 586/28 572/42 567/47
Total	**439/283**	**11.97±10.51**	0.6mm	**46208**	–	–/–

Table 1: Statistical Information and Data Distribution for two tasks in CIED Dataset.

Structural Anomaly Detection This task focuses on congenital anomalies including bony cochlear nerve canal stenosis, internal auditory canal stenosis, enlarged vestibular aqueduct, Mondini malformation, vestibular semicircular canal deformity, and Michel deformity. Since these anomalies are not mutually exclusive and highly imbalanced, we also provide a binary normal/abnormal label to support broader anomaly detection studies. The statistical details of the structural anomaly detection task are presented in Table 1.

Inner Ear Anatomical Structure Segmentation To provide spatial priors for the above tasks, we include an auxiliary segmentation task covering three clinically important structures: internal auditory meatus, cochlea, and cochlear nerve canal. Expert clinicians guided slice-wise polygonal annotations for volumetric structures and line-based keypoint annotation for canal diameter. These segmentation masks support integrating anatomical context into prediction and detection models.

4 Methodology

To effectively incorporate inner ear domain knowledge when adapting medical foundation models for inner ear analysis, we propose the Domain Knowledge Guided Tuning framework. In this section, we describe its key components in detail. We first present the overall pipeline in Section 4.1. Section 4.2 details the modeling of inner ear domain knowledge. Section 4.3 introduces the design of Domain Knowledge Aggregator, and Section 4.4 describes our adaptation for diverse architecture foundation models.

4.1 Pipeline Overview

In general, **DKGT** is a unified adaptation paradigm that can be applied to various model architectures. It integrates inner ear domain knowledge into foundation models by injecting features of key anatomical structures as domain-specific tokens.

An overview of the DKGT framework is shown in Figure 1. Guided by clinical experts in the inner ear domain, a domain knowledge generator is trained to detect three key anatomical structures and extract radiomic features that encapsulate both audiological and radiological knowledge. These features are encoded into learnable tokens by a domain knowledge aggregator, which integrates them with image features extracted by the foundation model across multiple network layers through a cross-attention–based interaction mechanism. The domain knowledge aggregator functions as a plug-and-play adaptation module, compatible with both ViT-based and CNN-based backbones, enabling effective knowledge injection without modifying pretrained weights.

4.2 Inner Ear Domain Knowledge Modeling

With guidance from domain experts, we identified three key anatomical structures of the inner ear that are critical for disease diagnosis and clinical evaluation: the internal auditory meatus, cochlea, and cochlear nerve canal. We then extracted geometric features that capture important characteristics of these structures, such as their shape and texture. These features serve as domain knowledge and are incorporated into the foundation model.

For relatively larger volumetric structures, such as the internal auditory meatus and cochlea, we extract radiomic features (Lambin et al. 2012) from the segmented regions, focusing on 3D shape descriptors. These include the surface area–to–volume ratio, sphericity, and 13 additional shape-related features. The shape features extracted from the internal auditory meatus and cochlea are denoted as F_1 and F_2, respectively. To characterize the cochlear nerve canal, we measure its width at each axial slice where it is visible. The width statistics of the cochlear nerve canal are then computed and used as the third set of features, denoted as F_3.

Subsequently, F_1, F_2, and F_3 are each projected to a unified feature dimension using three separate projection layers. This transformation enables the resulting domain knowledge tokens $\mathrm{t}_{i,j}$ to be seamlessly integrated into the intermediate layers of the foundation model, where they interact with the latent features extracted at each layer:

$$\mathrm{t}_{i,j} = \mathrm{ProjectionLayer}_{i,j}(\mathrm{F}_j) \tag{1}$$

where i represents the ith layer of foundation model, $j \in \{1, 2, 3\}$ represents the jth anatomical structure.

Segmentation and keypoint detection models were trained to extract these structures automatically.

4.3 Domain Knowledge Aggregator

Given expert domain knowledge represented by radiomic features extracted from key anatomical structures, we propose a Domain Knowledge Aggregator (DKA) to integrate

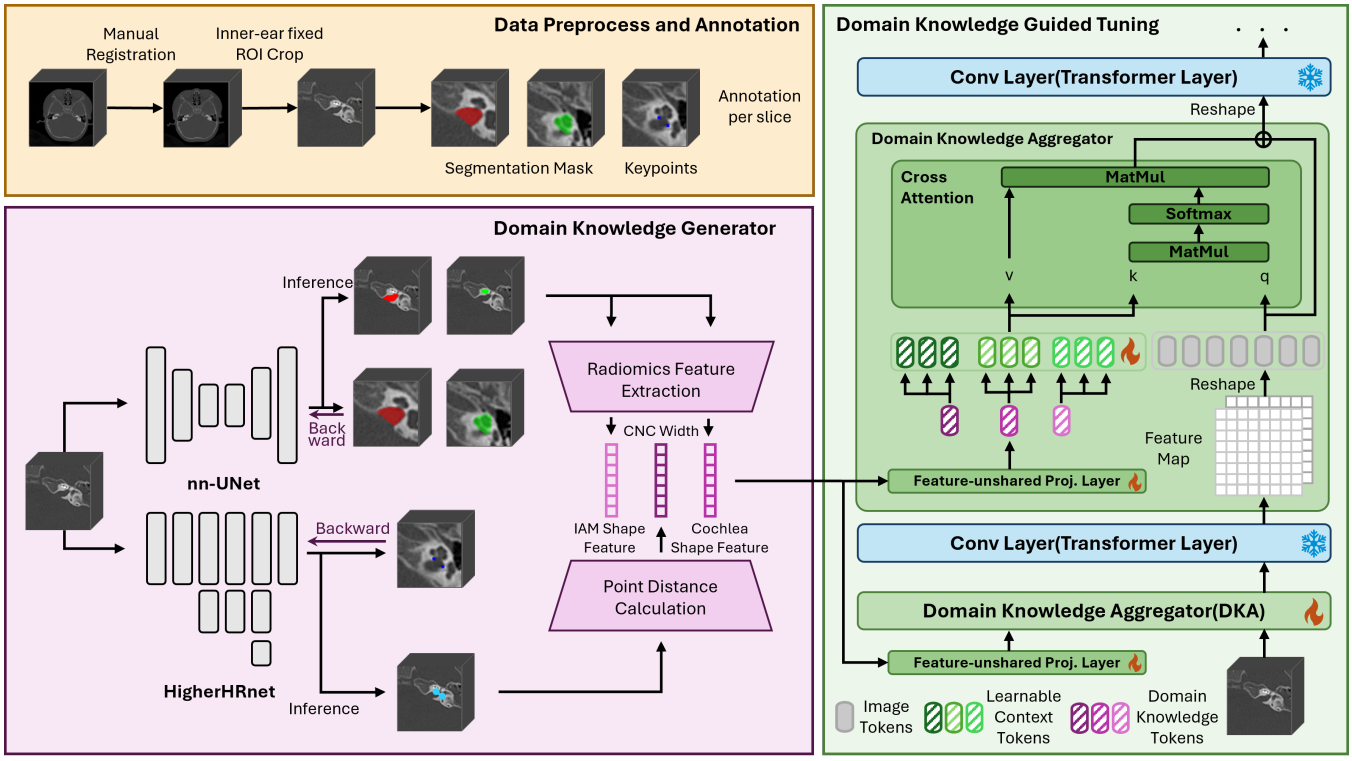

Figure 1: The overall architecture of our pipeline. DKGT consists of two main components: Domain Knowledge Generator and Domain Knowledge Aggregator. The generator extracts radiomic features from key anatomical structures and encodes them into domain knowledge tokens. These tokens are incorporated into multiple layers of the foundation model via DKA using a cross-attention mechanism. The plug-and-play design of DKGT is architecture-agnostic, supporting seamless integration with various backbone models.

this information into the medical foundation model. Specifically, DKA enables the set of domain knowledge tokens $\mathrm{t}_{i,j}$ encoded from radiomic features to interact with image features extracted at multiple intermediate layers of the foundation model via a cross-attention mechanism. By fine-tuning only the parameters of the DKA while keeping the foundation model frozen, we enable effective multi-layer interaction with domain knowledge while preserving the strong representational capacity of the pretrained foundation model.

To enable domain knowledge injection at a specific layer of the medical foundation model, we introduce a set of learnable context tokens with dimension C and length L for each layer, denoted as $\mathrm{p}_i \in \mathbb{R}^{L\times C}$ for the i-th layer. Then we divide the context tokens evenly into three groups and perform element-wise addition with each of the three domain knowledge tokens, which embed the domain knowledge into the context tokens as follows:

$$\mathrm{p}_i = \begin{bmatrix} \mathrm{p}_i[0:L/3] + \mathrm{t}_{i,1}, \\ \mathrm{p}_i[L/3:2L/3] + \mathrm{t}_{i,2}, \\ \mathrm{p}_i[2L/3:L] + \mathrm{t}_{i,3} \end{bmatrix} \tag{2}$$

DKA employs a cross-attention mechanism to enable interaction between the learnable context tokens and the latent features extracted by the foundation model layer. In this setup, the latent features act as queries to attend to the most relevant domain information encoded within the context tokens:

$$\mathrm{x}_i^* = softmax(\frac{\mathrm{q}_i \cdot \mathrm{k}_i^T}{\sqrt{C}})\mathrm{v}_i \tag{3}$$

where

$$\mathrm{q}_i = W_i^Q \mathrm{x}_i, \mathrm{k}_i = W_i^K \mathrm{p}_i, \mathrm{v}_i = W_i^V \mathrm{p}_i$$

where $\mathrm{x}_i \in \mathbb{R}^{DHW\times C}$ represents the flattened feature map of the $(i-1)$ th layer also the input of the i th layer, calculated using the query projection matrix $W_i^Q \in \mathbb{R}^{C\times C}$. The key vectors and value vectors are computed using learnable context tokens which are computed in Eq.2 and projection matrices $W_i^K \in \mathbb{R}^{C\times C}$ and $W_i^V \in \mathbb{R}^{C\times C}$.

Subsequently, we applied an out projection layer and employed a residual connection to update the image tokens, which can be formulated as:

$$\mathrm{x}_i = \mathrm{OutProjectionLayer}(\mathrm{x}_i^*) + \mathrm{x}_i \tag{4}$$

4.4 Unified Adaptation for Foundation Models

Thanks to its cross-attention–based design, where the original image features serve as queries, DKA preserves the shape of the intermediate latent features in the foundation

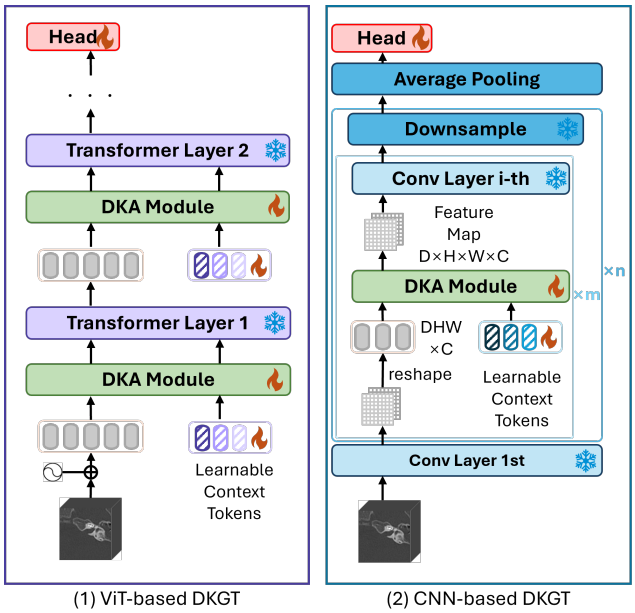

Figure 2: Unified adaptation of Domain Knowledge Aggregator(DKA) for ViT-based models and CNN-based models.

model. As a result, it functions as a unified adapter that can be seamlessly integrated into both ViT-based and CNN-based foundation models with minimal architectural modifications, as illustrated in Figure 2. For ViT-based models, DKA is inserted before each encoder layer and works regardless of classification head type. For CNN backbones, DKA is added before main convolutional stages for progressive refinement.

5 Experiment

5.1 Implementation Details

We select three representative foundation models—SAM-Med3D(Wang et al. 2024), CT-FM(Pai et al. 2025), and Merlin(Blankemeier et al. 2024)—for adaptation, as they differ in architecture and are pretrained on large-scale medical datasets, especially CT scans. SAM-Med3D uses a ViT-based encoder-decoder and is trained across diverse modalities, CT-FM adopts a U-Net-style CNN trained on a large CT corpus, while Merlin leverages a ResNet backbone with cross-modal supervision from CT images and radiology reports. These models offer strong generalization and make ideal candidates for inner ear CT analysis.

To meet the input requirements of SAM-Med3D, 3D volumes are resized from $128\times128\times64$ to $128\times128\times128$ via trilinear interpolation. All models are trained for 100 epochs with a batch size of 4 using the Adam optimizer (initial learning rate 1×10^{-4}, step decay every 20 epochs with factor 0.5) and early stopping. Data augmentation includes random horizontal flipping, arbitrary-angle rotation, and random translation. Identical hyperparameters are used across all methods. For the six-class classification task, focal loss is applied (γ=1.0, α proportional to class ratios) to address class imbalance; other tasks adopt cross-entropy loss. Performance is evaluated using 5-fold cross-validation, reporting mean and standard deviation.

Method	Total Tunable Params
full-finetuning - SM3D	98.76M(×1.000)
LoRA - SM3D(Zhang and Liu 2023)	9.95M(×0.101)
adapter - SM3D(Cheng et al. 2023)	655.23M(×6.634)
DKGT(ours) - SM3D	29.16M(×0.295)
full-finetuning - CT-FM	77.76M(×1.000)
DKGT(ours) - CT-FM	5.72M(×0.070)
full-finetuning - Merlin	121.88M(×1.000)
DKGT(ours) - Merlin	38.2M(×0.313)

Table 2: Comparison of total tunable parameters across different methods. SM3D denotes SAM-Med3D.

5.2 Comparative Experiment Results

To demonstrate the superiority of our proposed method, we compare it with several baselines, including an SVM using domain knowledge, two 3D CT classification methods (3DResNet(Hara, Kataoka, and Satoh 2018) and SuPreM(Li, Yuille, and Zhou 2025)), a late fusion of 3DResNet with domain knowledge, three medical foundation models with full fine-tuning, and several PEFT methods based on these models. Results in Table 3 lead to three findings: (1) Late fusion consistently outperforms plain 3DResNet, showing that domain knowledge introduces a useful inductive bias for inner ear tasks. (2) Medical foundation models, especially Merlin, perform better than task-specific models on the small inner ear dataset, highlighting their superior generalizability. (3) Our method outperforms all PEFT and full fine-tuning baselines by integrating DKA into multiple layers, effectively leveraging domain knowledge.

We also visualize attention maps of the DKA module (Figure 3). These show high attention scores in clinically important regions, including the internal auditory meatus, cochlea, and cochlear nerve canal, confirming that our model effectively focuses on inner ear structures.

Beyond explicitly modeled structures, the heatmaps highlight additional clinically relevant regions such as the geniculate ganglion, facial nerve, tensor tympani muscle, pyramidal eminence, and pharyngotympanic tube. This suggests that the model implicitly captures the facial nerve's relevance to cochlear implant outcomes and surgical complexity, offering valuable insights for preoperative imaging evaluation.

5.3 Ablation Study

We ablate different components of our DKGT method to thoroughly analyze their effects on outcome prediction and structural abnormality classification task.

Ablation of DKA Interaction Design We conduct ablation studies on the Domain Knowledge Aggregator (DKA) design. We evaluate: (1) removing the entire DKA module, (2) reversing the interaction where domain knowledge tokens query image tokens instead, and (3) two strategies

Task	Method	F1	ACC	AUC	SEN	PRE
Outcome Prediction	SVM	73.56±8.80	71.43±7.45	76.06±6.54	74.70±13.78	73.66±6.72
	3DResNet(Hara, Kataoka, and Satoh 2018)	74.17±5.86	66.62±7.99	60.81±13.55	88.18±12.44	65.24±7.29
	3DResNet - Late-fusion	78.33±4.65	70.39±7.88	59.37±11.99	96.67±4.08	66.26±6.91
	SuPreM(Li, Yuille, and Zhou 2025)	77.91±5.53	73.25±6.49	70.79±11.11	86.52±6.56	71.19±6.80
	adapter(Cheng et al. 2023) - SAM-Med3D	58.85±30.40	66.58±5.82	51.06±11.62	66.03±35.10	51.52±33.08
	LoRA(Zhang and Liu 2023) - SAM-Med3D	58.16±30.37	67.53±5.35	55.83±12.96	63.78±37.64	57.70±30.26
	Full-finetuning - SAM-Med3D	77.97±10.87	75.24±13.47	67.11±21.19	78.03±13.44	81.23±17.72
	Full-finetuning - CT-FM	75.71±6.87	70.52±9.97	71.20±7.88	84.39±14.19	70.80±9.30
	Full-finetuning - Merlin	77.17±6.95	75.06±7.31	73.64±10.81	79.24±16.55	79.23±10.39
	DKGT (ours) - SAM-Med3D	80.08±9.09	77.06±10.04	66.59±11.23	85.00±11.06	67.56±11.23
	DKGT (ours) - CT-FM	82.55±6.87	77.92±9.22	80.81±10.83	93.18±6.28	75.32±12.73
	DKGT (ours) - Merlin	**92.49±8.10**	**90.78±10.43**	**92.63±9.15**	**98.82±2.35**	**88.23±14.48**
Structural Abnormality Classification (binary)	SVM	73.51±6.74	82.89±4.28	88.48±4.00	65.77±6.61	83.49±7.62
	3DResNet(Hara, Kataoka, and Satoh 2018)	50.46±7.17	49.68±11.45	58.54±6.39	74.98±24.47	42.12±7.20
	3DResNet - Late-fusion	52.49±6.16	70.02±4.18	70.89±3.75	46.93±10.15	66.51±19.00
	SuPreM(Li, Yuille, and Zhou 2025)	67.50±6.99	79.97±3.78	80.72±6.11	58.45±5.07	80.73±12.65
	adapter(Cheng et al. 2023) - SAM-Med3D	35.57±8.45	56.80±12.84	61.03±4.91	39.40±30.75	50.41±18.35
	LoRA(Zhang and Liu 2023) - SAM-Med3D	30.21±13.93	68.73±3.11	64.89±4.74	20.36±10.69	79.55±11.92
	Full-finetuning - SAM-Med3D	50.97±7.13	63.95±16.03	68.81±5.96	53.04±24.50	62.08±17.15
	Full-finetuning - CT-FM	92.86±2.80	94.63±2.79	97.17±1.61	93.35±2.08	92.49±4.69
	Full-finetuning - Merlin	95.02±3.07	96.26±2.44	98.01±1.30	**95.42±3.77**	94.72±3.76
	DKGT (ours) - SAM-Med3D	76.71±4.99	84.05±4.90	85.38±4.99	71.75±4.69	82.75±7.71
	DKGT (ours) - CT-FM	93.30±2.05	95.12±1.70	97.38±1.03	93.50±2.43	93.19±3.23
	DKGT (ours) - Merlin	**95.90±1.28**	**97.07±1.10**	**98.51±0.63**	94.77±2.83	**97.19±2.81**
Structural Abnormality Classification (micro)	3DResNet(Hara, Kataoka, and Satoh 2018)	32.86±4.06	80.02±1.70	75.51±4.54	55.84±7.82	23.43±3.24
	3DResNet - Late-fusion	41.52±8.25	83.00±3.37	83.78±4.68	68.25±11.88	30.43±7.18
	SuPreM(Li, Yuille, and Zhou 2025)	52.05±3.37	87.70±1.39	89.78±2.34	75.86±4.27	39.86±4.36
	adapter(Cheng et al. 2023) - SAM-Med3D	29.80±3.62	76.19±1.03	72.47±2.47	57.47±4.57	20.27±3.24
	LoRA(Zhang and Liu 2023) - SAM-Med3D	30.21±13.93	68.73±3.11	64.89±4.74	20.36±10.69	**79.55±11.92**
	Full-finetuning - SAM-Med3D	37.39±2.50	81.81±2.78	80.44±3.69	61.76±8.09	27.10±2.68
	Full-finetuning - CT-FM	22.88±1.51	65.39±2.42	63.36±4.14	58.87±7.69	14.29±1.18
	Full-finetuning - Merlin	80.14±3.20	96.28±1.05	97.63±1.12	83.51±5.49	77.21±2.63
	DKGT (ours) - SAM-Med3D	45.33±4.32	84.55±1.95	86.70±1.61	72.58±5.31	33.04±3.78
	DKGT (ours) - CT-FM	48.33±3.64	85.08±2.87	87.71±2.70	78.45±5.88	35.00±3.12
	DKGT (ours) - Merlin	**80.79±2.68**	**96.31±0.88**	**97.65±0.86**	**86.93±4.53**	75.62±3.18

Table 3: The detailed results of different 3D CT advanced methods(%) for different tasks. Outcome Prediction represents the prediction of outcome after cochlear implantation surgery. Structural Abnormality Classification represents the detection of structural abnormalities in inner ear imaging. This includes determining the presence of abnormalities (binary classification) and identifying specific structural anomalies (six-class classification). Binary corresponds to the former, while micro corresponds to the latter.

for combining learnable context tokens with domain knowledge tokens: element-wise addition versus concatenation. The reversed interaction requires concatenating learnable tokens with image tokens as input to the frozen layer, which only works for ViT-based models due to their flexible input length. CNNs require fixed-size feature maps, making our DKA design applicable to both ViT and CNN architectures. Results in Table 4 show that DKA significantly improves SAM-Med3D's fine-tuning performance. The reversed interaction (domain knowledge tokens querying image tokens) performs slightly worse because domain knowledge tokens become less influential in the subsequent frozen transformer layer. For combination strategies, element-wise addition outperforms concatenation since it enables direct interaction between all learnable context tokens and domain knowledge tokens, while concatenation treats them independently.

Ablation of Different Domain Knowledge As mentioned above, we modeled three anatomical structures of the inner ear—the internal auditory meatus(IAM), the cochlea, and the bony cochlear nerve canal(CNC)—as domain knowledge. To evaluate their impact on model performance, we conducted experiments using different combinations of these domain knowledge on CT-FM and Merlin, which achieve better performance than SAM-Med3D. The results, summarized in Table 5, demonstrate the influence of different domain knowledge and their combinations on overall performance. Evidently, using domain knowledge of all three anatomical structures achieves the best performance.

Ablation of Segmentation Network Under clinical guidance, we annotated the internal auditory meatus and cochlea for each case. We used nn-UNet (3D)(Isensee et al. 2021) for segmentation and compared it with two 2D networks,

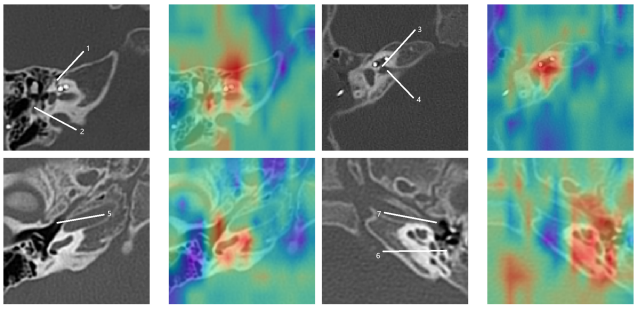

Figure 3: Visualization of attention map in DKA. Representative slices from the 3D volume are displayed, corresponding to the planes where the inner ear structures are prominently visible. The numbered labels in the figure correspond to the following anatomical structures: 1. geniculate ganglion, 2. vertical segment of the facial nerve, 3. bony cochlear nerve canal, 4. internal auditory canal, 5. tensor tympani muscle, 6. pyramidal eminence, and 7. tympanic orifice of the pharyngotympanic tube.

Task	Foundation Model	DKA Design				F1
		DKA	RevQ	Add	Concat	
OutPred	SAM-Med3D			✓		70.25±4.55
		✓	✓	✓		76.04±7.77
		✓			✓	78.28±9.45
		✓		✓		80.08±9.09
	CT-FM	✓			✓	80.91±4.10
		✓		✓		82.55±6.87
	Merlin	✓			✓	91.35±8.53
		✓		✓		**92.49±8.10**
StructAC	SAM-Med3D			✓		68.36±3.57
		✓	✓	✓		74.77±4.60
		✓			✓	75.87±5.19
		✓		✓		76.71±4.99
	CT-FM	✓			✓	92.36±2.25
		✓		✓		93.30±2.05
	Merlin	✓			✓	92.96±1.35
		✓		✓		**95.90±1.28**

Table 4: Ablation of different Domain Knowledge Aggregator design. 'DKA' indicates whether the DKA module is used. 'Reversed Query(RevQ)' denotes the design where domain knowledge tokens query image tokens (opposite to our approach). 'Add' represents element-wise addition between learnable context tokens and domain knowledge tokens, while 'Concat' denotes the concatenation of them.

Swin-Unet(Cao et al. 2022) and TransUNet(Chen et al. 2021). Dice scores were used for evaluation, and we further assessed the effect of different segmentation masks (Figure 4). nn-UNet (3D) achieved the best results, likely due to its ability to capture anatomical continuity across slices. While the choice of segmentation network had a minor effect on downstream performance, all segmentation-based approaches outperformed full fine-tuning. This confirms the effectiveness and robustness of using segmentation masks to generate domain knowledge.

Task	Foundation Model	Inner Ear Structure			F1	AUC
		IAM	cochlea	CNC		
OutPred	CT-FM				78.24±2.19	57.33±16.36
		✓	✓		79.98±8.68	74.76±16.01
		✓		✓	70.41±6.82	60.81±9.29
			✓	✓	82.28±4.67	78.18±12.16
		✓	✓	✓	**82.55±6.87**	**80.81±10.83**
	Merlin				91.35±8.53	**92.74±9.83**
		✓	✓		91.99±8.81	91.09±12.00
		✓		✓	91.72±9.23	91.56±10.52
			✓	✓	90.25±8.96	92.61±8.92
		✓	✓	✓	**92.49±8.10**	92.63±9.15
StructAC	CT-FM				89.99±2.52	96.95±1.54
		✓	✓		90.44±2.36	96.49±1.52
		✓		✓	91.28±1.83	97.14±0.78
			✓	✓	90.06±4.30	96.39±2.47
		✓	✓	✓	**93.30±2.05**	**97.38±1.03**
	Merlin				92.96±1.35	98.47±0.65
		✓	✓		94.06±1.78	98.53±0.67
		✓		✓	93.88±1.88	98.58±0.61
			✓	✓	94.10±1.56	**98.59±0.58**
		✓	✓	✓	**95.90±1.28**	98.51±0.63

Table 5: Ablation of different domain knowledge. IAM, cochlea, and CNC represent the internal auditory meatus, cochlea, and cochlear nerve canal respectively.

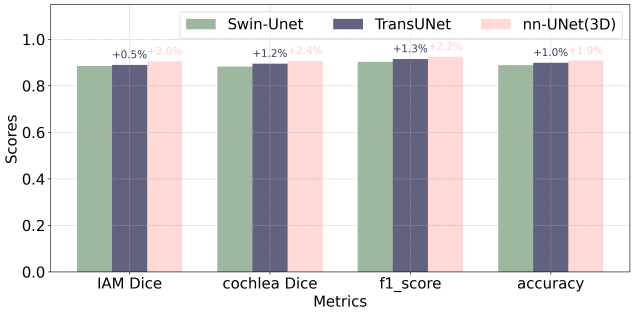

Figure 4: Comparison of different segmentation networks.

6 Conclusion

This work addresses data scarcity and domain knowledge integration in AI-based inner ear CT analysis through two contributions. We introduce the Children's Inner Ear CT Dataset (CIED), the first publicly available benchmark with 722 temporal bone CT scans supporting postoperative hearing recovery prediction, structural anomaly detection, and anatomical segmentation. We also propose Domain Knowledge Guided Tuning (DKGT), a parameter-efficient framework that incorporates clinical knowledge through a Domain Knowledge Aggregator. Experiments show that DKGT consistently outperforms existing methods on CIED, with ablations confirming the value of domain knowledge integration. This work provides a standardized dataset and generalizable framework to advance inner ear imaging analysis.

Acknowledgements

This work was supported by National Natural Science Fund of China (No.U25A20527, 62473286).

References

Blankemeier, L.; Cohen, J. P.; Kumar, A.; Van Veen, D.; Gardezi, S. J. S.; Paschali, M.; Chen, Z.; Delbrouck, J.-B.; Reis, E.; Truyts, C.; et al. 2024. Merlin: A vision language foundation model for 3d computed tomography. *Research Square*, rs–3.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218. Springer.

Casselman, J. W.; Offeciers, E. F.; De Foer, B.; Govaerts, P.; Kuhweide, R.; and Somers, T. 2001. CT and MR imaging of congential abnormalities of the inner ear and internal auditory canal. *European journal of radiology*, 40(2): 94–104.

Cen, J.; Fang, J.; Zhou, Z.; Yang, C.; Xie, L.; Zhang, X.; Shen, W.; and Tian, Q. 2025. Segment anything in 3d with radiance fields. *International Journal of Computer Vision*, 1–23.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; and Qiao, Y. 2022. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*.

Cheng, D.; Qin, Z.; Jiang, Z.; Zhang, S.; Lao, Q.; and Li, K. 2023. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*.

Fan, Q.; Huang, H.; Zhou, X.; and He, R. 2024. Lightweight vision transformer with bidirectional interaction. *Advances in Neural Information Processing Systems*, 36.

Fauser, J.; Stenin, I.; Bauer, M.; Hsu, W.-H.; Kristin, J.; Klenzner, T.; Schipper, J.; and Mukhopadhyay, A. 2019. Toward an automatic preoperative pipeline for image-guided temporal bone surgery. *International journal of computer assisted radiology and surgery*, 14: 967–976.

Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2023. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Girum, K. B.; Crehange, G.; and Lalande, A. 2021. Learning with context feedback loop for robust medical image segmentation. *IEEE transactions on medical imaging*, 40(6): 1542–1554.

Hara, K.; Kataoka, H.; and Satoh, Y. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 6546–6555.

Harnsberger, H.; Dart, D.; Parkin, J.; Smoker, W.; and Osborn, A. 1987. Cochlear implant candidates: assessment with CT and MR imaging. *Radiology*, 164(1): 53–57.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual Prompt Tuning. In *European Conference on Computer Vision (ECCV)*.

Ju, C.; Han, T.; Zheng, K.; Zhang, Y.; and Xie, W. 2022. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, 105–124. Springer.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; Van Stiphout, R. G.; Granton, P.; Zegers, C. M.; Gillies, R.; Boellard, R.; Dekker, A.; et al. 2012. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*, 48(4): 441–446.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Li, W.; Yuille, A.; and Zhou, Z. 2025. How well do supervised 3d models transfer to medical imaging tasks? *arXiv preprint arXiv:2501.11253*.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Liu, X.; Ji, K.; Fu, Y.; Tam, W. L.; Du, Z.; Yang, Z.; and Tang, J. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Lu, S.; Xie, J.; Wei, X.; Kong, Y.; Chen, B.; Chen, J.; Zhang, L.; Yang, M.; Xue, S.; Shi, Y.; et al. 2022. Machine learning-based prediction of the outcomes of cochlear implantation in patients with cochlear nerve deficiency and normal cochlea: a 2-year follow-up of 70 children. *Frontiers in Neuroscience*, 16: 895560.

Mao, Y.; Huang, K.; Guan, C.; Bao, G.; Mo, F.; and Xu, J. 2024. Dora: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. *arXiv preprint arXiv:2405.17357*.

Miyasaka, M.; Nosaka, S.; Morimoto, N.; Taiji, H.; and Masaki, H. 2010. CT and MR imaging for pediatric cochlear implantation: emphasis on the relationship between the cochlear nerve canal and the cochlear nerve. *Pediatric radiology*, 40(9): 1509–1516.

Neves, C.; Tran, E.; Kessler, I.; and Blevins, N. 2021. Fully automated preoperative segmentation of temporal bone structures from clinical CT scans. *Scientific reports*, 11(1): 116.

Pai, S.; Hadzic, I.; Bontempi, D.; Bressem, K.; Kann, B. H.; Fedorov, A.; Mak, R. H.; and Aerts, H. J. 2025. Vision foundation models for computed tomography. *arXiv preprint arXiv:2501.09001*.

Pfeiffer, J.; Rücklé, A.; Poth, C.; Kamath, A.; Vulić, I.; Ruder, S.; Cho, K.; and Gurevych, I. 2020. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*.

Rebuffi, S.-A.; Bilen, H.; and Vedaldi, A. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

Song, Z.; Gong, X.; Hu, G.; and Zhao, C. 2023. Deep perturbation learning: enhancing the network performance via image perturbations. In *International Conference on Machine Learning*, 32273–32287. PMLR.

Stickland, A. C.; and Murray, I. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *International Conference on Machine Learning*, 5986–5995. PMLR.

Vaidyanathan, A.; van der Lubbe, M. F.; Leijenaar, R. T.; van Hoof, M.; Zerka, F.; Miraglio, B.; Primakov, S.; Postma, A. A.; Bruintjes, T. D.; Bilderbeek, M. A.; et al. 2021. Deep learning for the fully automated segmentation of the inner ear on MRI. *Scientific reports*, 11(1): 2885.

Wang, H.; Guo, S.; Ye, J.; Deng, Z.; Cheng, J.; Li, T.; Chen, J.; Su, Y.; Huang, Z.; Shen, Y.; et al. 2024. Sam-med3d: towards general-purpose segmentation models for volumetric medical images. In *European Conference on Computer Vision*, 51–67. Springer.

Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.-S.; and Sun, M. 2024. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5: 30–38.

Yiin, R.; Tang, P.; and Tan, T. 2011. Review of congenital inner ear abnormalities on CT temporal bone. *The British journal of radiology*, 84(1005): 859–863.

You, E.; Lin, V.; Mijovic, T.; Eskander, A.; and Crowson, M. G. 2020. Artificial intelligence applications in otology: a state of the art review. *Otolaryngology–Head and Neck Surgery*, 163(6): 1123–1133.

Zhang, D.; Wang, J.; Noble, J. H.; and Dawant, B. M. 2020. HeadLocNet: Deep convolutional neural networks for accurate classification and multi-landmark localization of head CTs. *Medical image analysis*, 61: 101659.

Zhang, K.; and Liu, D. 2023. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, Y.; Shen, Z.; Zhao, Z.; Wang, S.; Wang, X.; Zhao, X.; Shen, D.; and Wang, Q. 2024. Melo: Low-rank adaptation is better than fine-tuning for medical image diagnosis. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.