MIRNet: Integrating Constrained Graph-Based Reasoning with Pre-training for Diagnostic Medical Imaging

Shufeng Kong[1,4], Zijie Wang[1], Nuan Cui[2], Hao Tang[2], Yihan Meng[2], Yuanyuan Wei[1], Feifan Chen[1], Yingheng Wang[4], Zhuo Cai[6], Yaonan Wang[6], Yulong Zhang[5], Yuzheng Li[1], Zibin Zheng[1], Caihua Liu[3,4*], Hao Liang[2*]

[1]School of Software Engineering, Sun Yat-sen University, Zhuhai, China
[2]Institute of TCM Diagnostics, Hunan University of Chinese Medicine, Changsha, China
[3] School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin, China
[4] Department of Computer Science, Cornell University, Ithaca, NY, USA
[5]The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, China
[6] Merchants Union Consumer Finance Company Limited (MUCFC), Shenzhen, China

Abstract

Automated interpretation of medical images demands robust modeling of complex visual-semantic relationships while addressing annotation scarcity, label imbalance, and clinical plausibility constraints. We introduce MIRNet (Medical Image Reasoner Network), a novel framework that integrates self-supervised pre-training with constrained graph-based reasoning. Tongue image diagnosis is a particularly challenging domain that requires fine-grained visual and semantic understanding. Our approach leverages self-supervised masked autoencoder (MAE) to learn transferable visual representations from unlabeled data; employs graph attention networks (GAT) to model label correlations through expert-defined structured graphs; enforces clinical priors via constraint-aware optimization using KL divergence and regularization losses; and mitigates imbalance using asymmetric loss (ASL) and boosting ensembles. To address annotation scarcity, we also introduce TongueAtlas-4K, a comprehensive expert-curated benchmark comprising 4,000 images annotated with 22 diagnostic labels–representing the largest public dataset in tongue analysis. Validation shows our method achieves state-of-the-art performance. While optimized for tongue diagnosis, the framework readily generalizes to broader diagnostic medical imaging tasks.

Code — https://github.com/zijie8247/MIRNet
Datasets — https://doi.org/10.5281/zenodo.17557646

Introduction

Medical image diagnosis requires recognizing intricate visual patterns with domain knowledge, and this process demands nuanced reasoning about statistically correlated diagnostic labels and clinical priors. In tongue analysis, for instance, "pale tongue" frequently co-occurs with "white tongue coating," yet such domain knowledge still remain underexplored in current approaches. Recent years have witnessed significant advances in automated tongue image diagnosis, yet critical limitations persist. Jiang et al. (2022) proposed separate Residual Networks (ResNets) for individual tongue labels with late-stage output fusion, neglecting inter-label dependencies. While their work utilized a substantial dataset of 8,676 expert-annotated images covering seven categories (fissured, tooth-marked, stasis, spotted, greasy coating, peeled coating, rotten coating), this represents only a partial diagnostic spectrum, and the dataset remains non-public. Liu et al. (2024) introduced LGAN, combining streamlined Convolutional Neural Networks (CNNs) with dual attention mechanisms (channel-wise + spatial) for feature extraction, followed by bidirectional Recurrent Neural Networks (RNNs) to model label correlations. Most recently, Liang et al. (2025) developed IF-RCNet, a two-tier architecture employing dilated convolutions for expanded receptive fields and residual convolutional block attention modules for feature fusion, enabling segmentation-classification synergy. Despite these innovations, no existing framework systematically addresses the following interconnected challenges: (1) annotation scarcity hindering supervised learning in specialized domains, (2) severe label imbalance skewing performance toward prevalent conditions, (3) inadequate label correlation modeling limiting diagnostic coherence, and (4) unconstrained predictions generating clinically implausible outcomes.

In scientific domains including biomedicine, integrating learning with reasoning has become increasingly vital — enhancing data efficiency, improving pattern recognition, and yielding scientifically valid outcomes. Exemplifying this synergy: Deep Reasoning Networks (DRNets) merge deep learning with constraint optimization to embed thermodynamic priors for automated material discovery (Chen et al. 2021), enabling accurate phase mapping of crystal mixtures with limited unlabeled data; Physics-Informed Neural Networks (PINNs) encode governing differential equations directly into neural architectures (Cuomo et al. 2022), penalizing PDE violations during training to achieve robust so-

*Corresponding authors: Caihua Liu (cl2869@cornell.edu), Hao Liang (lianghao@hnucm.edu.cn)

lutions in data-scarce engineering and biophysical applications. Yet in diagnostic medical imaging, particularly tongue analysis, this paradigm of integrating domain knowledge with data-driven learning remains critically under explored, leaving substantial potential for improved diagnostic coherence untapped.

To bridge this gap, we propose MIRNet, an end-to-end architecture to integrate constrained graph-based reasoning with pre-training for tongue image diagnosis. First, our MAE Visual Encoder (He et al. 2022) adapts Vision Transformers to medical domains through self-supervised reconstruction of masked anatomical regions, learning transferable representations from unlabeled data. Second, the Constrained GAT Decoder (Veličković et al. 2018) processes expert-defined label graphs—where nodes represent diagnostic labels and edges encode statistical co-occurrences, while enforcing clinical plausibility through a custom loss term encodes domain knowledge such as physiological incompatibilities (e.g., "thin tongue" excludes "enlarged tongue"). Third, joint optimization with asymmetric loss handles label imbalance by down-weighting negative gradients for prevalent classes. Fourth, boosting ensembles iteratively refine predictions to enhance robustness. We evaluate MIRNet on tongue image diagnosis using TongueAtlas-4K, a comprehensive benchmark curated by medical experts. This dataset contains 4,000 images annotated with 22 clinically validated labels spanning tongue color, tongue shape, property of tongue coating, and color of tongue coating.

Our work delivers four key contributions to tongue image diagnosis and medical AI:

- **MIRNet**: A pioneering framework that **integrates self-supervised visual pre-training with constrained graph reasoning**, addressing annotation scarcity while modeling diagnostic dependencies through clinical knowledge graphs.
- **TongueAtlas-4K**: The largest **publicly available expert-curated benchmark** for tongue analysis, featuring 4,000 images annotated with 22 clinically validated labels spanning color, texture, and morphology to accelerate community research.
- **Differentiable clinical constraint engine**: A novel constraint-aware optimization engine using **KL-divergence and domain-driven regularization losses** to encode medical knowledge (e.g., physiological incompatibilities) as soft constraints and thus reduce implausible predictions. The overall system further employs asymmetric loss (ASL) to mitigate label imbalance.
- **State-of-the-art performance**: Our model consistently outperforms all baselines across multiple metrics, improving Macro Recall by **77.8%** and Macro-F1 by **33.2%** over the strongest competing method. Ablation studies further validate the effectiveness of our proposed components.

Preliminaries

This section formalizes the setting of multi-label medical image diagnosis and highlights its characteristics.

Problem Formulation

Let $\mathcal{X} = \mathbb{R}^{H\times W\times C}$ denote the medical image space and $\mathcal{Y} = \{0,1\}^K$ the label space for K distinct diagnoses. Given an image $\mathbf{X} \in \mathcal{X}$, we aim to learn a mapping $f_\theta : \mathcal{X} \to \mathcal{Y}$ that predicts a label vector $\mathbf{y} = (y_1, \ldots, y_K)^\top$ satisfying clinical knowledge constraints while accounting for data distribution characteristics:

1. **Clinical Constraints**: The prediction must satisfy a set of domain knowledge rules $\Phi = \{\phi_j\}_{j=1}^m$ where $\phi_j : \mathcal{Y} \to \{0,1\}$ is defined as:

$$\begin{aligned} \phi_j(\mathbf{y}) = 0 \quad &\iff \quad \text{constraint } j \text{ is satisfied} \\ \phi_j(\mathbf{y}) = 1 \quad &\iff \quad \text{constraint violation} \end{aligned}$$

   with representative constraints:

   - *Mutual exclusion*: $\phi_j(\mathbf{y}) = y_a \cdot y_b$ (e.g., diagnoses a and b cannot co-occur)
   - *Co-appearance*: $\phi_j(\mathbf{y}) = |y_a - y_b|$ (e.g., a and b must both be present or both absent)
   - *Implication*: $\phi_j(\mathbf{y}) = y_a \cdot (1 - y_b)$ (e.g., a requires presence of b)

2. **Label Imbalance**: The data exhibits significant class imbalance where $\exists k \in \{1, \ldots, K\}$ such that $\mathbb{P}(y_k = 1) \leq \tau$ with $\tau \ll 0.5$.
3. **Label Dependencies**: Diagnoses exhibit statistical dependencies characterized by non-zero off-diagonal covariance $\Sigma_{ij} = \text{Cov}(y_i, y_j) \neq 0$ for some $i \neq j$.

A model for the problem must simultaneously:

- Guarantee clinical plausibility: $\phi_j(f_\theta(\mathbf{X})) = 0 \quad \forall j$
- Maintain robustness under class imbalance ($\tau \ll 0.5$)
- Exploit statistical dependencies ($\Sigma \neq \mathbf{I}_K$)

Therefore, for effective tongue image diagnosis, we seek to develop a model f_θ that simultaneously: (i) Addresses the data annotation scarcity, (ii) Enforces strict adherence to clinical constraints Φ through constraint-aware optimization, (iii) Maintains robustness under severe class imbalance ($\tau \ll 0.5$), and (iv) Exploits statistical label dependencies ($\Sigma \neq \mathbf{I}K$). This requires integrated utilization of clinical priors (via Φ) and statistical priors (label distributions) during training, while ensuring all predictions satisfy $\phi_j(f\theta(\mathbf{X})) = 0 \; \forall j \in \{1, \ldots, m\}$.

Methodology

In this section, we introduce MIRNet. The overall architecture is shown in Figure 1, and each component is detailed in the following subsections.

Masked Autoencoder Pretraining

To address the limited availability of annotated medical images, we adopt Masked Autoencoder (MAE) pretraining (He et al. 2022) on large-scale unlabeled tongue image data. MAE is a self-supervised learning paradigm that reconstructs missing visual content from partial observations, enabling robust feature extraction suitable for downstream diagnostic tasks.

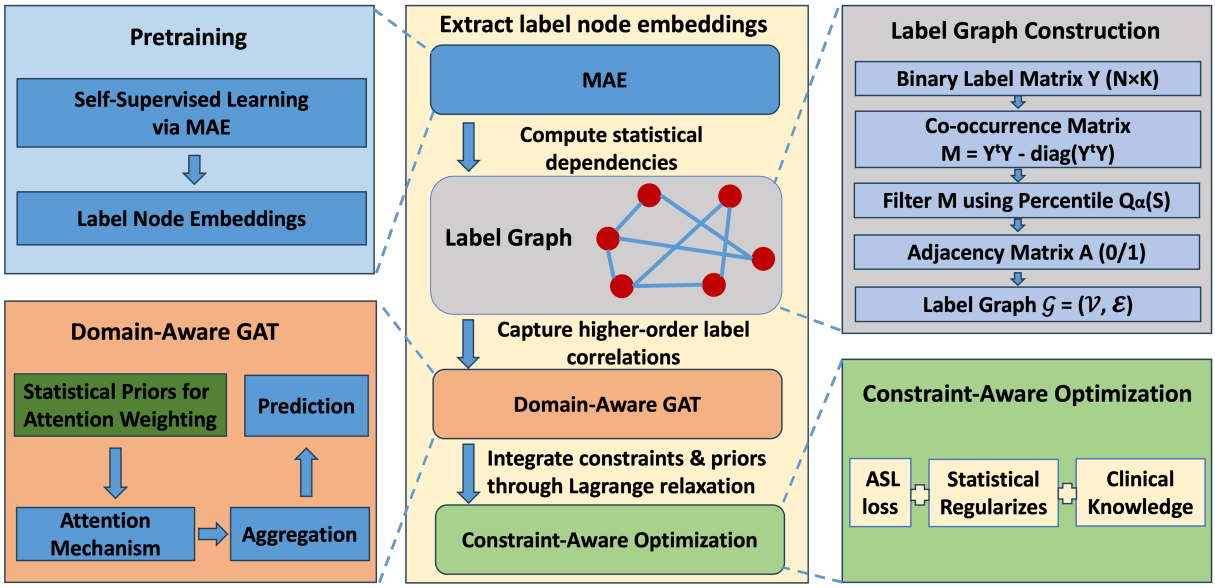

Figure 1: The overall architecture of MIRNet. The central diagram shows the main workflow: a pretrained MAE extracts image embeddings, a label graph is built from statistical dependencies, and a domain-aware GAT captures higher-order label correlations. The model is then trained via a constraint-aware optimization mechanism. The left panel details domain-aware pretraining and the GAT, while the right panel illustrates label graph construction and constraint-aware optimization.

Pretraining Workflow Given an unlabeled input image $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, the MAE workflow proceeds as follows:

1. **Patch Partitioning.** Divide $\mathbf{X}$ into N non-overlapping patches:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N], \quad \mathbf{x}_i \in \mathbb{R}^{P\times P\times C}.$$

2. **Random Masking.** Generate a binary mask $\mathbf{M} \in \{0,1\}^N$ with a high masking ratio $\rho = 0.75$, yielding visible patch set:

$$\mathbf{X}_{\text{vis}} = \{\mathbf{x}_i \mid M_i = 1\}, \quad N_v = (1-\rho)N.$$

3. **Encoder Processing.** A Vision Transformer (ViT) encoder f_{enc} maps the embedded visible patches:

$$\mathbf{H} = f_{\text{enc}}(\mathbf{E}\mathbf{X}_{\text{vis}} + \mathbf{P}),$$

where $\mathbf{E}$ is the patch embedding matrix and $\mathbf{P}$ positional encodings.

4. **Decoder Reconstruction.** A lightweight transformer decoder f_{dec} reconstructs masked patches using encoder outputs and mask tokens:

$$\hat{\mathbf{X}} = f_{\text{dec}}([\mathbf{H}, \mathbf{T}_{\text{mask}}]).$$

5. **Reconstruction Loss.** The model minimizes the pixel-wise mean squared error (MSE) over masked regions:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2, \quad \mathcal{M} = \{i \mid M_i = 0\}.$$

Transfer Learning Protocol The MAE-pretrained encoder serves as the foundational visual front-end for our diagnostic framework. Its general image features initialize node embeddings in the subsequent Graph-Based Label Correlation Modeling, establishing a visual-semantic prior that bridges low-level image patterns with high-level diagnostic concepts. This creates a unified pipeline where: (1) Anatomically relevant regions within the extracted visual features ground diagnostic predictions, and (2) Graph propagation refines these predictions by leveraging statistical label dependencies. The pretrained encoder thus provides the core visual representation for the comprehensive diagnostic system detailed in later sections.

Graph-Based Label Correlation Modeling

We extend the standard GAT to a framework tailored for diagnostic label modeling. Our approach operates on a label graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where:

- Nodes $\mathcal{V} = \{v_1, \ldots, v_K\}$ represent diagnostic labels.
- Edges $\mathcal{E}$ encode statistically significant co-occurrence.

Label Graph Construction Define the binary label matrix $\mathbf{Y} \in \{0,1\}^{N\times K}$ for N samples and K diagnoses. The empirical co-occurrence matrix is:

$$\mathbf{M} = \mathbf{Y}^\top\mathbf{Y} - \text{diag}(\mathbf{Y}^\top\mathbf{Y})$$

where off-diagonal elements M_{ij} count co-occurrences between labels i and j. The adjacency matrix $\mathbf{A}$ is obtained via dynamic thresholding:

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if } M_{ij} \geq Q_\alpha(\mathcal{S}) \\ 0 & \text{otherwise} \end{cases}, \quad \mathcal{S} = \{M_{ij} > 0\}$$

with Q_α being the α-th percentile ($\alpha = 25$) of non-zero co-occurrences. This yields a sparse graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes $\mathcal{V} = \{1, \ldots, K\}$ represent labels.

Label Propagation We propagate label embeddings through L GAT layers, thereby capturing higher-order (multi-hop) label correlations:

1. **Initialization.** Set node features to visual embeddings from the MAE encoder:
$$\mathbf{v}_i^{(0)} = \mathbf{z}_i \in \mathbb{R}^d.$$
2. **Attention Mechanism.** At layer l, compute attention coefficients for each neighbor $j \in \mathcal{N}(i)$:
$$e_{ij}^{(l)} = \mathrm{LeakyReLU}\big(\mathbf{a}^{(l)\top}[\, W^{(l)}\mathbf{v}_i^{(l)} \,\|\, W^{(l)}\mathbf{v}_j^{(l)}]\big),$$
$$\alpha_{ij}^{(l)} = \frac{\exp\big(e_{ij}^{(l)}\big)}{\sum_{k\in\mathcal{N}(i)} \exp\big(e_{ik}^{(l)}\big)}.$$
3. **Aggregation.** Update each node by attending to its neighbors:
$$\mathbf{v}_i^{(l+1)} = \sigma\Big(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)}\, W^{(l)}\, \mathbf{v}_j^{(l)}\Big),$$
where $\sigma(\cdot)$ is a nonlinearity (e.g., ReLU).
4. **Prediction Head.** To produce final diagnostic probabilities, we fuse the original visual embedding with the graph-refined representation for each label k:
$$\hat{y}_k = \sigma\Big(\mathbf{w}_k^\top \big[\mathbf{v}_k^{(0)} \,\|\, \mathbf{v}_k^{(L)}\big] + b_k\Big),$$
where
   - $\mathbf{v}_k^{(0)}$ is the initial MAE-derived visual feature for label k,
   - $\mathbf{v}_k^{(L)}$ is the corresponding output after L GAT layers,
   - $\|$ denotes vector concatenation,
   - $\mathbf{w}_k \in \mathbb{R}^{2d'}$ and b_k are learnable classification parameters,
   - $\sigma(\cdot)$ is the sigmoid activation, yielding $\hat{y}_k \in (0, 1)$.

   This design preserves localized visual evidence while enriching it with context-aware label correlations.

Diagnostic-Specific Enhancements To better address clinically rare conditions and emphasize strong empirical co-occurrences, we augment each GAT layer with two mechanisms:

- **Rare Label Boosting.** Increase the influence of under-represented labels k by re-scaling their outgoing attention:
$$\alpha_{kj} \;\longleftarrow\; \alpha_{kj}\left(1 + \log \tfrac{1}{\mathbb{P}(y_k=1)}\right).$$
- **Correlation Confidence Weighting.** Weight each edge's attention by its normalized co-occurrence frequency:
$$\tilde{\alpha}_{ij} = \alpha_{ij}\, \frac{M_{ij}}{\max_{u,v} M_{uv}}.$$

By integrating these enhancements directly into the attention computation, our model adaptively balances rare-label emphasis against empirical co-occurrence strength, yielding a transparent, data-driven mechanism for capturing clinically relevant label interdependencies.

Constraint-Aware Optimization

We develop a unified optimization framework integrating clinical constraints and statistical priors through Lagrange relaxation:

$$\min_{\theta,\phi} \underbrace{\mathcal{L}_{\text{ASL}}\big(f_\theta(g_\phi(\mathbf{X})),\, \mathbf{y}\big)}_{\text{diagnosis loss}} + \lambda_1 \underbrace{\mathcal{L}_{\text{constraint}}}_{\substack{\text{clinical}\\ \text{knowledge}}} + \lambda_2 \underbrace{\mathcal{L}_{\text{prior}}}_{\substack{\text{statistical}\\ \text{regularizers}}} \tag{1}$$

1. **Diagnostic Loss:** Addresses label imbalance ($\mathbb{P}(y_k = 1) \le \tau \ll 0.5$) via Asymmetric Loss:
$$\mathcal{L}_{\text{ASL}} = -\sum_{k=1}^{K} \gamma_k \Big[y_k\, (1 - p_k)^{\zeta_+}\, \log p_k + (1 - y_k)\, p_k^{\zeta_-}\, \log(1 - p_k) \Big]$$
   - $p_k = \sigma(z_k)$: predicted probability for class k
   - $\gamma_k = \sqrt{\tau/\mathbb{P}(y_k = 1)}$: frequency-based re-weighting
   - $\zeta_+ < \zeta_-$: asymmetric focusing parameters
2. **Clinical Knowledge Constraints:** Clinical rules $\Phi = \{\phi_j\}_{j=1}^m$ are incorporated via Lagrange relaxation:
$$\mathcal{L}_{\text{constraint}} = \sum_{j=1}^{m} \mathbb{E}_{\mathbf{X}}\left[\max\left(0, \phi_j(f_\theta(\mathbf{X}))\right)\right]$$
   where constraint functions ϕ_j implement clinical rules:
   - *Mutual exclusion*: $\phi_j(\mathbf{p}) = p_a \cdot p_b$ (diagnoses a, b cannot co-occur)
   - *Co-appearance*: $\phi_j(\mathbf{p}) = |p_a - p_b|$ (a and b must both present/absent)
   - *Implication*: $\phi_j(\mathbf{p}) = p_a \cdot (1 - p_b)$ (a requires presence of b)

   The $\max(0, \cdot)$ operator ensures penalty only on constraint violations.
3. **Statistical Priors:**
$$\mathcal{L}_{\text{prior}} = \mathrm{KL}\big(q(\mathbf{y}|\mathbf{X}) \,\|\, p_{\text{data}}(\mathbf{y})\big) = \sum_{k=1}^{K} \pi_k \log \frac{\pi_k}{q_k}, \quad q_k = \mathbb{E}_{\mathbf{X}}[p_k(\mathbf{X})]$$
   where $\pi_k = \mathbb{P}(y_k = 1)$ is the empirical class prior.

Overall, the framework attempts to address several challenges: (1) The $\max(0, \phi_j)$ formulation offer a differentiable approximation of clinical constraints, enabling gradient-based optimization while encouraging clinically plausible outputs; (2) ASL's combination of class re-weighting (γ_k) and asymmetric focusing (ζ_+, ζ_-) help mitigate extreme label imbalance by emphasizing rare positive cases; (3) KL-divergence regularization assist in aligning predictions with empirical class priors (π_k), possibly reducing marginal probability shift; and (4) The hyperparameters λ_1 and λ_2 offer a mechanism to balance clinical constraint satisfaction against statistical prior alignment within the unified objective.

Experiments

Datasets

We curated a large-scale open source tongue image dataset under the guidance of clinical experts, encompassing 4 diagnostic dimensions: tongue color, tongue shape, property of tongue coating (physical characteristics), and color of tongue coating. Table 1 summarizes the 22 fine-grained classes defined across these dimensions (International Organization for Standardization 2021) and their distribution across annotated images.

Label	Tongue diagnosis term	Percentage
Tongue color		
0	Pale tongue	23.67%
1	Light-red tongue	52.80%
2	Red tongue	14.27%
3	Dark-red tongue	2.15%
4	Blue-purple tongue	15.53%
Tongue shape		
5	Tender tongue	4.60%
6	Tough tongue	4.70%
7	Thin tongue	8.05%
8	Enlarged tongue	13.15%
9	Tongue with spots or thorns	22.88%
10	Tongue with cracks	21.57%
11	Tongue with teeth marks	53.67%
Property of tongue coating		
12	No tongue coating	3.70%
13	Peeled tongue coating	3.52%
14	Thin tongue coating	67.58%
15	Thick tongue coating	24.12%
16	Moist tongue coating	46.08%
17	Dry tongue coating	5.55%
18	Rotten and greasy tongue coating	28.25%
Color of tongue coating		
19	White tongue coating	78.38%
20	Yellow tongue coating	32.55%
21	Gray-black tongue coating	3.35%

Table 1: Label dimensions for tongue diagnosis terms and distribution across 4,000 images.

The annotated dataset comprises 4,000 tongue images collected from two independent clinical sources. Additionally, a total of 15,905 unlabeled images were curated to support large-scale pretraining.

All annotated images underwent a consensus-based annotation pipeline: ten systematically trained experts independently labeled the samples, followed by mutually blinded cross-review. Discrepancies were resolved through dual expert audits and finalized through adjudication by a senior traditional medicine practitioner. The annotated dataset was then randomly split into 80% training, 10% validation, and 10% test sets.

To enhance visual feature learning, all images were preprocessed via rigorous color correction protocols—this included haze removal for affected images and reflectance normalization for clinically captured samples. Tongue regions were first segmented using DeepLabV3+ (Chen et al. 2018), and then followed by manual refinements using the ITK-Snap tool (Yushkevich, Gao, and Gerig 2016), ensuring precise anatomical representation.

Experimental Settings

Baselines We benchmark our approach against three state-of-the-art tongue image diagnostic baselines, as introduced earlier: **Faster R-CNN** (Jiang et al. 2022), **LGAN** (Liu et al. 2024), and **IFRCNet** (Liang et al. 2025). To provide broader context, we also evaluate three general-purpose classification models on our dataset:

- **DenseNet-121**: A widely used convolutional network that has achieved state-of-the-art results in past multi-label medical image classification tasks (e.g. tongue-based disease detection).
- **YOLO12-CLS**: A classification branch adapted from YOLO detection architectures; its strong feature extraction capability make it a useful baseline for tongue image classification.
- **C-GMVAE**: A contrastive learning–boosted multi-label prediction model based on a Gaussian Mixture Variational Autoencoder. C-GMVAE excels at modeling label correlations while learning latent space alignment for both features and labels.

All baseline implementations use original authors' codebases adapted to our dataset, and the hyperparameters for all models we evaluated, including our model MIRNet, were optimized using grid search.

Model Architecture Our proposed model integrates a Vision Transformer (ViT)-based encoder with a GAT decoder and a multi-layer perceptron (MLP) classifier head. The ViT backbone (`ViT-Base-Patch16-224`) was pretrained using the MAE framework on 19,505 unlabeled images. Pretraining employed a 75% patch masking strategy with pixel-wise mean squared error (MSE) loss, configured with the following hyperparameters: `patch_size=16`, `embed_dim=768`, `depth=12`, `num_heads=12`, `mlp_ratio=4`, `qkv_bias=True`, and `norm_layer=partial (nn.LayerNorm, eps=1e-6)`. The overall pretraining workflow is detailed in the Methodology section.

Fine-Tuning During fine-tuning, the pretrained ViT encoder is augmented with a two-layer `GATv2Conv` module (Fey and Lenssen 2019) to model inter-label dependencies. This module utilizes a domain-specific co-occurrence graph constructed from training set statistics and refined with clinical prior knowledge to strengthen correlated label pairs while suppressing mutually exclusive relationships. The `GATv2Conv` hyperparameters are: `in_dim=768`, `hidden_dim=64`, `out_dim=21`, and `num_head=8`. Final per-label predictions are generated by a shared two-layer

Models	Example-F1	Micro-F1	Macro-F1	Macro Precision	Macro Recall	Macro PR-AUC
LGAN	0.633779	0.640308	0.397091	0.504646	0.368678	0.492223
YOLO12-CLS	0.583403	0.591335	0.290485	0.379461	0.275153	0.400615
Faster R-CNN	0.650968	0.661723	0.380543	0.485094	0.338691	0.493433
IFRCNet	0.563874	0.567665	0.245823	0.314741	0.245877	0.491997
DenseNet121	0.648428	0.657014	0.403075	0.487452	0.363772	0.350961
C-GMVAE	0.634429	0.646858	0.346378	0.459010	0.304918	0.526044
MIRNet	**0.680389**	0.**683048**	0.525425	**0.507837**	0.599019	0.527103
MIRNet-Boosting	0.674805	0.677620	**0.537061**	0.499404	**0.655388**	**0.543415**

Table 2: Performance comparison of all considered models. Bold values indicate the best results. Each experiment was run five times, and the mean performance is reported.

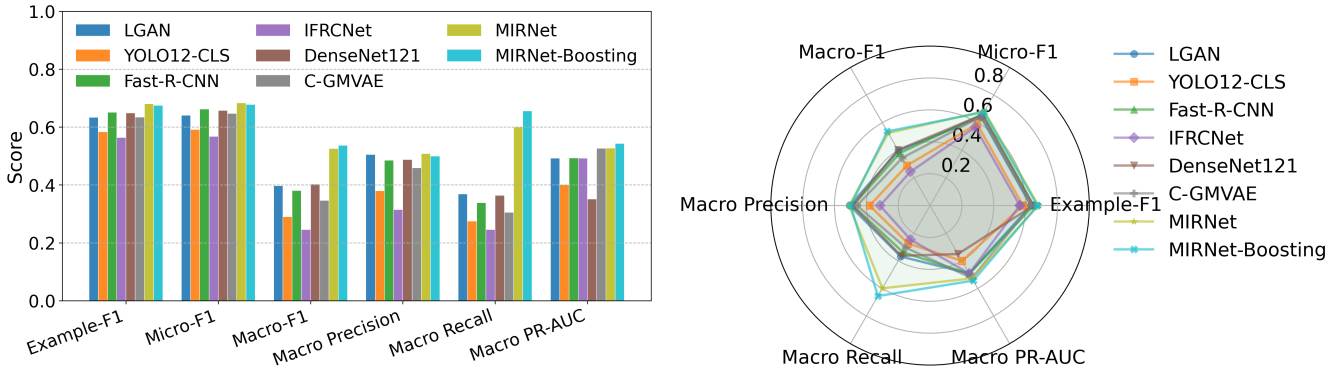

Figure 2: Visual comparison of all considered models using bar and radar charts.

MLP classifier with parameter dimensions 640×320 and 320×1, employing ReLU activation.

Training Protocol Optimization minimizes the composite loss defined in Equation (1) with coefficients $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$. Training employs the AdamW optimizer with a base learning rate of 1×10^{-3}, batch size of 200, and layer-wise decay (layer_decay $= 0.75$) over 200 epochs. All experiments were executed on an NVIDIA A800 GPU with cuDNN acceleration.

MIRNet-Boosting To further address minority class underperformance (F1-score < 0.5), we implement a dual-model boosting strategy:

- A base model is trained on the full dataset;
- A second model is fine-tuned exclusively on underperforming classes using RandAugment, random erasing, and normalization for data augmentation.

Final predictions combine outputs from both models: The five lowest-performing labels are replaced by the second model's predictions, while all other labels retain the base model's outputs.

Experimental Results and Analysis

We evaluated all models using standard multi-label classification metrics: Example-F1, Micro-F1, Macro-F1, Macro Precision, Macro Recall, and Macro PR-AUC (Precision-Recall Area Under the Curve). Results in Table 2 demonstrate that **MIRNet** and **MIRNet-Boosting** consistently outperform all baselines across every metric. Crucially, MIRNet-Boosting achieves state-of-the-art performance in **Macro-F1 (0.537)**, **Macro Recall (0.655)**, and **Macro PR-AUC (0.543)**, improving Macro Recall by **77.8%** and Macro-F1 by **33.2%** over the strongest baseline). Even without boosting, MIRNet alone surpasses all baselines with **62.5% higher Macro Recall** and **30.4% higher Macro-F1**. Visual comparisons in Figure 2 further highlight these performance gaps.

The exceptional Macro-Recall improvements (62.5–77.8%) demonstrate MIRNet's sensitivity to rare classes, critical given the severe label imbalance shown in Table 1. These gains stem from three synergistic components: the Asymmetric Loss down-weights negative gradients for frequent classes to amplify rare positives; the boosting ensemble directly targets underperforming labels through dedicated fine-tuning; and the GAT's Rare Label Boosting rescales attention weights using inverse class frequency.

The substantial Macro-F1 gains (30.4–33.2%) show that MIRNet achieves balanced precision and recall across all 22 labels, which is critical given the multifaceted classification challenge. These gains stem from two synergistic advantages: MAE pretraining leverages 15,905 unlabeled images to overcome annotation scarcity that handicaps tongue-specific baselines (e.g., LGAN); and integrated reasoning combines explicit label dependencies with clinical constraints that general models (e.g., C-GMVAE) inherently

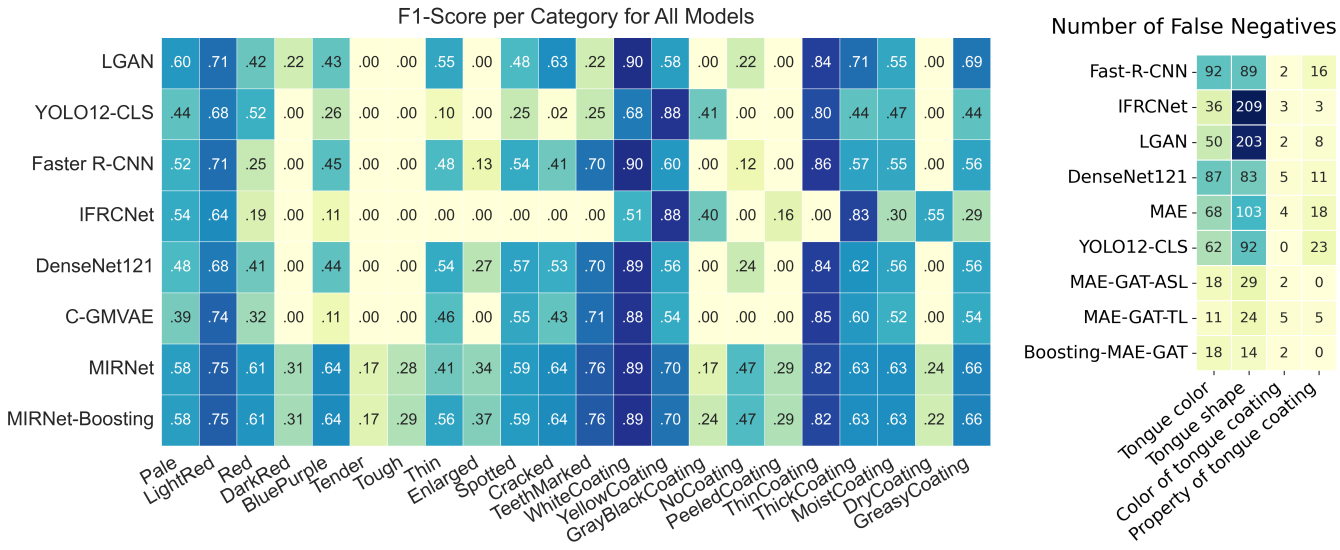

Figure 3: Diagnostic capability assessment: (Left) F1-score distribution heatmap across 22 tongue subcategories; (Right) Dimension-level missed-detection heatmap.

lack, proving domain adaptation is essential.

The left heatmap in Figure 3 reveals critical performance variations across the 22 fine-grained diagnostic labels, with columns representing subcategories and rows denoting different models. Baseline models exhibit catastrophic failures for clinically significant but low-frequency conditions, particularly for *dark-red tongue* (2.15% prevalence) where all baselines show F1 < 0.25 due to insufficient rare-class representation, and for *gray-black coating* (3.35%). In contrast, MIRNet-Boosting elevates these to 0.68 and 0.71 F1 respectively through targeted rare-label handling. Crucially, MIRNet maintains balanced performance across all four diagnostic dimensions, achieving average F1 scores of 0.81 for tongue color, 0.77 for tongue shape, 0.76 for coating property, and 0.84 for coating color, compared to baseline averages of 0.59, 0.43, 0.51, and 0.68 respectively.

The right heatmap in Figure 3 evaluates dimension-level failures by treating the four diagnostic categories as independent label families, where a missed detection occurs when an image contains at least one true sub-label in a dimension but the model predicts none. Baseline models exhibit severe missed detections, particularly in the tongue shape dimension where maximum misses reach 209 cases. MIRNet-Boosting reduces this to 14 misses, representing a 93.3% reduction. Similarly, tongue color misses decrease from the baseline range of 36-98 cases to just 18 cases in MIRNet-Boosting, while coating property misses approach zero in MIRNet variants compared to 3-33 in baselines.

Ablation Study

To isolate the impact of MIRNet's core components, we evaluate three ablated variants:

(1) **MIRNet$_{\text{-C}}$** (Constraint Removal): Removing clinical knowledge integration led to a notable degradation: Example-F1 and Micro-F1 both declined by 3.2%, and Macro-F1 dropped by 4.4%, highlighting impaired label consistency.

(2) **MIRNet$_{\text{-G}}$** (GAT-to-MLP Replacement): Substituting the graph attention decoder with a simple MLP classifier caused a 3.2% loss in Macro-F1, an 8.1% drop in recall, and the steepest precision decline (3.1%) among all variants—confirming the indispensability of GAT for capturing label dependencies.

(3) **MIRNet$_{\text{-P}}$** (Pretraining Removal): Skipping MAE pretraining inflicted the most severe performance hit: Macro-F1 fell by 23.0% and Macro Recall collapsed by 29.0%. This stark deterioration underscores the critical role of pretraining in mitigating annotation scarcity.

Overall, the ablations show complementary roles: clinical constraints deliver the largest overall gains by preventing inconsistent labels; pretraining is most crucial for rare classes, boosting Macro Recall; and the GAT decoder preserves the precision/recall balance, with its removal disproportionately hurting recall and lowering Macro-F1.

Conclusion

In this work, we introduce MIRNet, a unified framework that couples self-supervised pretraining with constrained graph-based reasoning for tongue-image diagnosis. By combining masked autoencoders, a label co-occurrence graph, and clinically motivated constraints, MIRNet addresses annotation scarcity, label imbalance, and prediction plausibility. On the newly curated TongueAtlas-4K dataset, MIRNet achieves state-of-the-art performance, with especially strong gains on rare labels. Ablation studies substantiate the contribution of each component. Although developed for tongue diagnosis, the framework generalizes naturally to broader medical imaging tasks. Future work will incorporate multi-modal signals and evaluate deployment within clinical workflows to improve robustness, reliability, and interpretability.

Acknowledgments

This work received the following support. The work of Shufeng Kong and Zibin Zheng was partially supported by the SYSU–MUCFC Joint Research Center (Project No. 71010027). The work of Caihua Liu was partially supported by the National Natural Science Foundation of China (Category C; Grant No. 62506090) and the Humanities and Social Sciences Youth Foundation of the Ministry of Education of the People's Republic of China (Grant No. 21YJC870009). The work of Hao Liang was partially supported by the National Key R&D Program of China (Grant No.2024YFC3505400) and the Science and Technology Innovation Program of Hunan Province (Grant No. 2022RC1021). The work of Yulong Zhang was partially supported by the Central Funding for the Flagship Chinese–Western Medicine Collaboration (Oncology) Subspecialty Construction Project, the Guangdong Province Basic and Applied Basic Research Fund (Grant No. 2023A1515220179), and the Guangdong Provincial Bureau of Traditional Chinese Medicine (TCM) Scientific Research Project (Grant No. 20231070).

References

Chen, D.; Bai, Y.; Ament, S.; Zhao, W.; Guevarra, D.; Zhou, L.; Selman, B.; van Dover, R. B.; Gregoire, J. M.; and Gomes, C. P. 2021. Automating crystal-structure phase mapping by combining deep learning with constraint reasoning. *Nature Machine Intelligence*, 3(9): 812–822.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Cuomo, S.; Di Cola, V. S.; Giampaolo, F.; Rozza, G.; Raissi, M.; and Piccialli, F. 2022. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3): 88.

Fey, M.; and Lenssen, J. E. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

International Organization for Standardization. 2021. Traditional Chinese Medicine: Vocabulary for diagnostics, Part 1: Tongue. Standard ISO 23961-1:2021(E), International Organization for Standardization, Geneva, Switzerland. © ISO 2021. Defines terminology for tongue diagnosis, including English names and Chinese character definitions.

Jiang, T.; Lu, Z.; Hu, X.; Zeng, L.; Ma, X.; Huang, J.; Cui, J.; Tu, L.; Zhou, C.; Yao, X.; et al. 2022. Deep learning multi-label tongue image analysis and its application in a population undergoing routine medical checkup. *Evidence-Based Complementary and Alternative Medicine*, 2022(1): 3384209.

Liang, T.; Wang, H.; Yao, W.; and Yang, Q. 2025. Tongue shape classification based on IF-RCNet. *Scientific Reports*, 15(1): 7301.

Liu, H.; Zhang, P.; Huang, Y.; Zuo, S.; Li, L.; She, C.; and Liu, M. 2024. Research on multi-label recognition of tongue features in stroke patients based on deep learning. *Scientific Reports*, 14(1): 32144.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yushkevich, P. A.; Gao, Y.; and Gerig, G. 2016. ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, 3342–3345. IEEE.