DCMM-Transformer: Degree-Corrected Mixed-Membership Attention for Medical Imaging

Huimin Cheng[1*], Xiaowei Yu[2*], Shushan Wu[3], Luyang Fang[3], Chao Cao[4], Jing Zhang[4], Tianming Liu[5], Dajiang Zhu[4], Wenxuan Zhong[3], Ping Ma[3†]

[1]Department of Biostatistics, Boston University
[2]Department of Computer Science, Missouri University of Science and Technology
[3]Department of Statistics, University of Georgia
[4]Department of Computer Science and Engineering, University of Texas at Arlington, USA
[5]School of Computing, University of Georgia, USA
pingma@uga.edu

Abstract

Medical images exhibit latent anatomical groupings, such as organs, tissues, and pathological regions, that standard Vision Transformers (ViTs) fail to exploit. While recent work like SBM-Transformer attempts to incorporate such structures through stochastic binary masking, they suffer from non-differentiability, training instability, and the inability to model complex community structure. We present DCMM-Transformer, a novel ViT architecture for medical image analysis that incorporates a Degree-Corrected Mixed-Membership (DCMM) model as an additive bias in self-attention. Unlike prior approaches that rely on multiplicative masking and binary sampling, our method introduces community structure and degree heterogeneity in a fully differentiable and interpretable manner. Comprehensive experiments across diverse medical imaging datasets, including brain, chest, breast, and ocular modalities, demonstrate the superior performance and generalizability of the proposed approach. Furthermore, the learned group structure and structured attention modulation substantially enhance interpretability by yielding attention maps that are anatomically meaningful and semantically coherent.

Introduction

Medical image analysis is essential for disease diagnosis, prognosis, and informing clinical decisions (Varoquaux and Cheplygina 2022; Yu et al. 2023c, 2024a,b). Vision Transformers (ViTs) and their variants, such as TransUNet (Chen et al. 2024) and Swin Transformer (Liu et al. 2021), have achieved state-of-the-art performance. These models segment images into small, fixed-size patches (tokens) and employ self-attention mechanisms to model complex, long-range relationships across different anatomical structures. By capturing such dependencies, ViTs have demonstrated impressive results across a range of tasks, including image classification, segmentation, and object detection (Hatamizadeh et al. 2022; Yu et al. 2023a).

Challenges for Medical Images. Despite rapid advances, medical image analysis remains challenging and often necessitates architectural designs tailored to the unique properties of medical data (Huang et al. 2025, 2024; Yu et al. 2022a; Litjens et al. 2017; Yu et al. 2022b; Zhang et al. 2025; Yu et al. 2025). Medical images are spatially organized into distinct anatomical regions, such as tumor areas, healthy tissue, and blood vessels, that form meaningful clusters within the image (Esteva et al. 2021; Yu et al. 2021; Chavoshnejad et al. 2021). Patches from the same anatomical or pathological region tend to be correlated, sharing features that are particularly informative for clinical tasks. Conversely, patches from different regions (such as tumor versus healthy tissue, or across distinct organs) often exhibit distinct underlying biological properties, resulting in weaker relationships between them.

Motivation for Community-Aware Attention. This community characteristic motivates the development of community-aware attention mechanisms, which encourage the model to focus on interactions within anatomically or pathologically coherent regions (Wang et al. 2019). Such approaches are particularly valuable for detecting subtle patterns: in tumor detection, for example, the signal from any single patch within a tumor may be weak or ambiguous. By aggregating information across patches within the same community, the model can amplify these weak signals, resulting in a more coherent and accurate diagnostic assessment (Ilse, Tomczak, and Welling 2018; Yu et al. 2023b).

Limitations of Standard ViTs. Nevertheless, conventional transformer-based attention mechanisms do not take this underlying community structure into account. Traditional attention approaches treat tokens uniformly, calculating weights solely based on learned feature similarities without considering anatomical groupings (Vaswani et al. 2017; Dosovitskiy et al. 2021). As a result, the model may sometimes focus on regions that are distant or unrelated in a clinical sense, simply because they appear similar in feature space. This lack of anatomical awareness can cause the model to overlook clinically important regions, become more vulnerable to noise or artifacts, and make its predic-

[*]These authors contributed equally.
[†]Corresponding author.

tions harder to interpret (Li et al. 2023b). These limitations underscore the necessity for transformers that explicitly incorporate community structure (Li et al. 2023a).

Limitations of SBM-Transformer. Recent work has explored incorporating explicit community structures into attention mechanisms. One notable example is SBM-Transformer (Cho et al. 2022) which integrates a mixed-membership Stochastic Block Model (SBM) into each attention head to learn latent communities among image patches. To reduce computational costs, the SBM-Transformer employs binary attention masks that randomly sample token connections based on probabilities learned from the SBM. Only token pairs selected by these masks can attend to each other, resulting in sparse attention. However, this binary masking introduces non-differentiability, which requires surrogate gradient methods like the Straight-Through Estimator and leads to optimization bias and convergence instability (Cho et al. 2022) . Additionally, collapsing probabilistic connection strengths into hard, binary decisions discards useful information about the strength of token relationships and increases variance during both training and inference. A further limitation is that the SBM-Transformer ignores degree heterogeneity. In practice, image patches exhibit significant variation in connectivity and importance. For example, tumor cores often act as local hubs by interacting with many other regions, an effect not captured under the assumption of uniform connectivity.

Proposed Method. To overcome these limitations, we propose DCMM-Transformer, a method integrating a Degree-Corrected Mixed-Membership (DCMM) model into ViTs. An overview of DCMM-Transformer is shown in Figure 1. In particular, we model the latent community structure among image patches using the DCMM framework, which captures both overlapping community membership and degree heterogeneity. The DCMM module generates a learned probability matrix, where each entry reflects the likelihood of interaction between a pair of patches, informed by their shared community memberships and individual connectivity. Our approach adds the learned probability matrix directly to attention logits as an additive bias, bypassing random sampling and binary discretization. This community-aware bias helps the model aggregate weak or distributed signals from related patches, such as those within the same tumor region, thereby enhancing the detection of clinically important patterns that might be missed when treating all patches equally.

Advantages. DCMM-Transformer offers three key advantages: (1) Structural guidance without distortion: Unlike the SBM-Transformer (Cho et al. 2022), which uses multiplicative masking to impose sparsity and risks distorting semantic relationships by over-suppressing weak connections or unevenly amplifying others, our additive bias offers structural guidance while preserving data-driven insights. (2) Training stability: Addition maintains stable gradients and normalization, avoiding vanishing gradient issues. (3) Enhanced modeling: DCMM captures both degree heterogeneity and mixed membership, capabilities absent in standard SBMs. We provide extensive evaluations across five medical image classification tasks, demonstrating an average 3.7% improvement in performance over standard ViT and its variants, as well as enhanced interpretability for clinical applications. Ablation studies further reveal that excluding degree heterogeneity reduces performance by 2.3%.

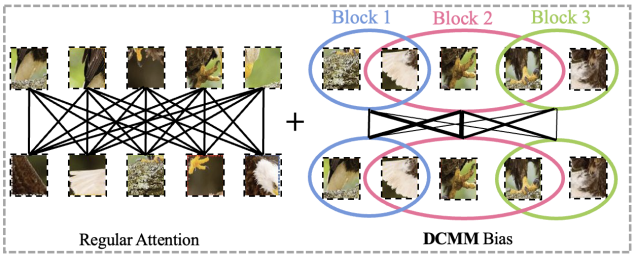

Figure 1: Overview of the DCMM-Transformer. The DCMM model imposes latent community structure among image patches, grouping them into blocks (e.g., Blocks 1–3 in blue, red, and green), and assigns probabilities to the interactions within and between these communities. This community-based bias is then added to the standard attention logits.

Related Work

In this section, we present two areas of related work: SBMs and Vision Transformers.

Related Work on SBM

SBM. The stochastic block model (Holland, Laskey, and Leinhardt 1983; Cheng et al. 2018; Liu, Cheng, and Zhang 2019) is a fundamental probabilistic model for networks with community structure. In the SBM, each node is assigned to one of L latent communities (or blocks), and the probability of an edge between any two nodes depends only on their community memberships. Let $A \in (0,1)^{n\times n}$ denote the adjacency matrix of the network, and let $z_i \in 1,\dots,L$ be the community label of node i. The block matrix $B \in (0,1)^{L\times L}$ specifies the probability of connection between communities, so that edges are sampled independently according to

$$\Pr(A_{ij}=1) = B_{z_i z_j} \tag{1}$$

where B_{z_i,z_j} is the probability of an edge between node i and node j. While the SBM provides a simple and effective framework for modeling community structure, it assumes that all nodes within a block have identical expected degrees, which is often violated in real-world networks exhibiting heterogeneous degree distributions.

Degree-Corrected SBM. To address the limitation of the homogeneous degree assumption in SBM, the Degree-Corrected Stochastic Block Model (DC-SBM) (Karrer and Newman 2011) introduces node-specific degree parameters. Under DC-SBM, the probability of an edge between nodes i and j depends not only on their community memberships, but also on individual propensity parameters $\theta_i > 0$ and $\theta_j > 0$ that control expected node degrees, allowing for more realistic modeling of networks with heterogeneous degrees:

$$\Pr(A_{ij}=1) = \theta_i\,\theta_j\,B_{z_i z_j} \tag{2}$$

Mixed-Membership SBM. The Mixed Membership SBM (MMSBM) (Airoldi et al. 2008) extends the classical SBM to handle overlapping communities by replacing hard community assignments with soft membership vectors. Each node i has a probability membership vector: $\boldsymbol{\pi}_i = (\pi_i(1), \ldots, \pi_i(L))^\top \in [0,1]^L$, where $\sum_{\ell=1}^{L} \pi_i(\ell) = 1$. The probability of an edge between nodes i and j is

$$\Pr(A_{ij} = 1) = \sum_{\ell=1}^{L} \sum_{\ell'=1}^{L} \pi_i(\ell)\, \pi_j(\ell')\, B_{\ell\ell'} = \boldsymbol{\pi}_i^\top B \boldsymbol{\pi}_j \quad (3)$$

When each $\boldsymbol{\pi}_i$ is a standard basis vector (i.e., nodes belong to only one community), Equation (3) reduces to (1) in classical SBM.

Degree-Corrected Mixed-Membership SBM (DCMM). The most general model combines mixed membership and degree heterogeneity (Jin, Ke, and Luo 2024):

$$\Pr(A_{ij} = 1) = \theta_i\, \theta_j\, \boldsymbol{\pi}_i^\top B \boldsymbol{\pi}_j \quad (4)$$

Previous models are special cases of DCMM: setting $\theta_i = 1$ recovers the MMSBM, while taking both $\theta_i = 1$ and one-hot $\boldsymbol{\pi}_i$ yields the classical SBM.

Related Work on Transformer

Vision Transformer and Attention. ViTs process images by treating them as sequences of tokens, similar to how language models handle text. First, the input image is divided into n non-overlapping, flattened 2D patches (tokens), resulting in a matrix $X \in \mathbb{R}^{n \times d}$, where d is the dimension of each patch embedding. The core of ViT is the self-attention mechanism, which enables the model to capture dependencies and interactions between all pairs of patches. Second, within each attention head, the model computes pairwise relationships among tokens by projecting them into query (Q), key (K), and value (V) representations using learnable weight matrices $Q = XW^Q, \quad K = XW^K, \quad V = XW^V$, where $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_h}$ and d_h is the head dimension. Attention scores are then computed as the scaled dot product $\frac{QK^\top}{\sqrt{d_h}}$, followed by row-wise softmax normalization to obtain weights $\sigma\left(\frac{QK^\top}{\sqrt{d_h}}\right)$ by softmax. The output is finally derived as a weighted sum of values:

$$\text{Attn}(Q, K, V) = \sigma\left(\frac{QK^\top}{\sqrt{d_h}}\right) V$$

This process allows each patch to attend to all other patches based on their learned similarities, enabling the model to capture long-range dependencies across the entire image.

SBM-Transformer. SBM-Transformer (Cho et al. 2022) directly replaces the dense attention mechanism with a sparse, learnable alternative guided by an MMSBM, with the main purpose of reducing the computational burden of the Transformer. In particular, the connection probability p_{ij} between tokens i and j is given by Equation (3), where π and $\mathbf{B}$ are parameters learned from neural networks. A binary attention mask M_{ij} is then sampled as $M_{ij} \sim \text{Bernoulli}(p_{ij})$. This mask is applied to the attention mechanism via masked attention:

$$\text{Attn}_{\text{mask}}(Q, K, V, M) = \sigma_M\left(M \odot \frac{QK^\top}{\sqrt{d_h}}\right) V,$$

where $\odot$ denotes element-wise multiplication, and $\sigma_M(\cdot)$ is the masked softmax that normalizes only over the non-zero entries in each row of M.

Even though SBM-Transformer reduces computational cost, it suffers from these limitations: (1) Non-differentiability and optimization instability: The binary sampling is non-differentiable, as the sampling function lacks a continuous derivative with respect to p_{ij}. This breaks gradient flow, requiring surrogate methods like Straight-Through Estimator, but this surrogate introduces bias and high variance. (2) Information loss through hard thresholding: Binarization discards the probabilistic information in p_{ij}, which can encode subtle relationships. For example, $p_{ij} = 0.99$ and $p_{ij} = 0.1$ might both yield $M_{ij} = 1$, but their connection strengths differ significantly. (3) Lack of degree heterogeneity in the MMSBM: MMSBM fails to capture degree heterogeneity, a characteristic of real networks, including image patch graphs where hub patches (e.g., tumor cores) connect broadly, while peripherals do not.

Transformer with Additive Bias. Another related line of research enhances attention mechanisms with explicit bias terms that encode spatial or structural relationships. For example, ACC-ViT (Ibtehaz et al. 2024) introduces Atrous Attention with multiple dilation rates: $\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + b\right) V$, where b denotes branch-specific relative positional biases, and outputs are fused through gated aggregation. The Swin Transformer (Liu et al. 2021) restricts attention to local patch blocks while adding a learnable relative position bias for substantial gains. Similarly, MaxViT (Tu et al. 2022) employs multi-axis self-attention to outperform standard mechanisms. However, these methods do not take into account the community structure.

In sum, while prior works have shown the promise of community-aware attention and additive bias, they remain limited, e.g., SBM-based masking is non-differentiable, while standard bias terms do not consider the latent community structure.

DCMM-Transformer

To address these limitations, the DCMM-Transformer augments the standard transformer attention logits with a connection probability matrix under DCMM, allowing the model to learn both soft community assignments and node-specific centrality in a fully differentiable manner.

Overview of DCMM-Transformer

The key innovation of DCMM-Transformer is employing a novel self-attention logits by adding a learned community structure matrix $P \in \mathbb{R}^{n \times n}$ which is generated under the DCMM:

$$\text{Attn}_{\text{DCMM}}(Q, K, V, P) = \sigma\left(\frac{QK^\top}{\sqrt{d_h}} + \lambda P\right) V \quad (5)$$

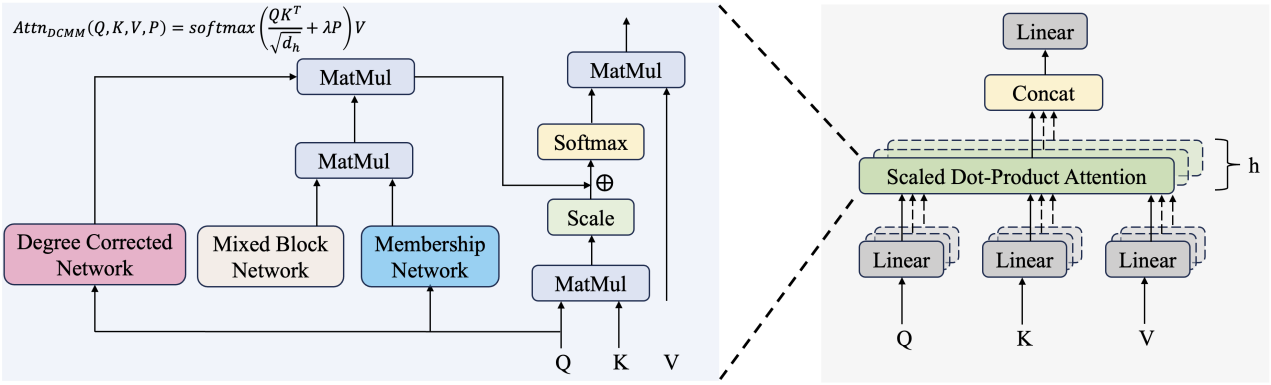

Figure 2: Overview of the proposed DCMM-Transformer mechanism integrated into the self-attention module of Vision Transformers. For each token, a soft group membership vector and a degree scalar are constructed from the query representation. These are used to construct a structured attention bias based on learnable inter-group affinities. The bias is added to the standard attention scores to modulate them before softmax.

where λ is a tunable hyperparameter controlling the contribution of the community structure. When $\lambda = 0$, the model reduces to a standard transformer; as λ increases, the community structure dominates.

Figure 2 illustrates the DCMM-Transformer. The DCMM bias P is constructed on the left side of the diagram through a series of specialized networks and operations applied primarily to Q. Specifically, each token (image patch) i is processed by two parallel neural networks: The membership network module computes soft community membership vectors π_i for each patch, indicating its association with each latent community. The degree corrected network module produces a node-specific degree parameter θ_i, capturing the relative centrality or importance of each patch. The mixed block network module introduces L trainable cluster embeddings for latent communities.

Rationale of additive bias. We choose additive bias over multiplicative bias in Equation (5) due to the following reasons. (1) **Gentle Enhancement vs. Exponential Scaling**. When the DCMM bias P is added to the logits, the effect after the softmax is a multiplication by $\exp(\lambda P_{ij})$, where P_{ij} is always between 0 and 1. This means the community prior can gently increase the corresponding attention weights by at most a factor of e^{λ} (when $P_{ij} = 1$), or leave them nearly unchanged (when P_{ij} is close to 0), without erasing the information from the original feature similarities. In contrast, with multiplicative bias, the logits are directly scaled by P_{ij} before exponentiation, so any value of P_{ij} less than 1 can drastically shrink the effective logit, leading to a nearly uniform or uninformative attention pattern and causing the model to overlook meaningful relationships. (2) **Training Stability.** Let $l = \frac{QK^\top}{\sqrt{d_h}}$. For additive bias, the softmax gradient is: $\frac{\partial \sigma(l+\lambda P)}{\partial l} = \text{diag}(\sigma(l+\lambda P)) - \sigma(l+\lambda P)\sigma(l+\lambda P)^\top$. This standard softmax Jacobian remains well-conditioned regardless of bias values λP, since the gradient structure is unchanged—only the softmax outputs shift. However, for multiplicative bias $P \odot l$, the gradient becomes: $\frac{\partial \sigma(P \odot l)}{\partial l} = \text{diag}(P) \cdot [\text{diag}(\sigma(P \odot l)) - \sigma(P \odot l)\sigma(P \odot l)^\top]$ Since the $\text{diag}(P)$ scales each gradient component by the corresponding $P_{ij} \in [0, 1]$, this can lead to a vanishing gradient problem for low P_{ij} which is very common in sparse networks.

Procedures of DCMM-Transformer

We detail below how DCMM-Transformer constructs and updates the community structure matrix P within the model pipeline.

(1) Cluster Embedding Initialization: To represent latent communities, we introduce L trainable cluster embeddings $C \in \mathbb{R}^{L \times d_l}$. These cluster embeddings parameterize the features of each community within the attention head's representation space.

(2) Block Connecting Probability: To model the probability of interaction between communities, we construct a non-negative block connection probability matrix $B \in (0, 1)^{L \times L}$. Each entry B_{ij} represents the probability of connection between community i and community j. The diagonal entries of B indicate intra-community connection probabilities, while the off-diagonal entries represent inter-community probabilities. We obtain B by applying a row-wise softmax to the inner product $CC^\top$, where C is the community embeddings.

(3) Soft Community Memberships: For each token i, we then construct its soft community membership vector $\pi_i \in [0, 1]^L$, where each element π_{ij} represents how strongly token i belongs to community j. Specifically, π_i is obtained by

$$\tilde{q}_i = \text{MLP}(q_i), \quad \boldsymbol{\pi}_i = \text{Sigmoid}(\tilde{q}_i C^\top) \tag{6}$$

where MLP is a two-layer MLPs with ReLU activations that project queries to the community space.

(4) Node-Specific Degree Correction: To account for heterogeneous degree distribution of tokens, we introduce a node-specific degree parameter for each token,

$$\theta_i = \text{Sigmoid}(W_\theta q_i)$$

where $W_\theta \in \mathbb{R}^{1 \times d}$ are learned parameters.

Method	ChestXray	SIIMACR	INbreast	ADNI	EyeDisease	Average
ViT	92.3	86.9	79.7	66.7	92.5	83.6
SBM-Transformer	95.8	85.8	78.4	64.2	91.3	83.1
DeiT	95.8	86.8	83.8	65.4	91.7	84.5
ConViT	93.6	79.6	81.1	71.6	92.4	83.7
Swin Transformer	95.0	87.6	84.8	65.4	92.9	85.1
MaxViT	94.5	86.4	83.3	71.7	93.2	85.8
DCMM-Transformer	**96.0**	**88.2**	**85.1**	**74.1**	**93.3**	**87.3**

Table 1: Comparative classification accuracy (%) of our DCMM-Transformer and baseline Vision Transformer variants across five medical imaging datasets. Bold values indicate the highest accuracy for each dataset.

(5) Constructing the Connecting Probability: Combining all learned components, each entry of the community structure matrix is:

$$p_{ij} \;=\; \theta_i \theta_j \, \boldsymbol{\pi}_i^{\top} B \, \boldsymbol{\pi}_j \tag{7}$$

With P constructed, the DCMMM multi-head self-attention (MSA_{DCMM}) is:

$$\mathrm{MSA}_{DCMM}(Q, K, V, P) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W^O$$

where each attention head is defined as $\mathrm{head}_i = \mathrm{Attn}_{\mathrm{DCMM}}(QW_i^Q, KW_i^K, VW_i^V, P)$. The learnable parameter matrices W_i^Q, W_i^K, W_i^V are the projections in each subspace (head), and W^O is the projection for all subspaces. Multi-head attention helps the model to jointly aggregate information from different representation subspaces at various positions.

Learning Objective and Loss Function. To enhance interpretability and encourage the model to learn distinct and meaningful community assignments for each token, we add an entropy regularization term on the soft community membership vectors. Specifically, the entropy loss $\mathcal{L}_{\text{entro}}$ penalizes uniform (high-entropy) memberships and promotes sharper, more confident assignments, where each token clearly belongs to a specific set of communities. Mathematically, the entropy of community assignment loss is:

$$\mathcal{L}_{\text{entro}} = -\frac{1}{n}\left[\sum_{i=1}^{n}\sum_{j=1}^{L} \pi_{ij} \log\left(\max(\pi_{ij}, \epsilon)\right)\right] \tag{8}$$

where ϵ is a small constant to prevent taking the log of zero. The total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha \, \mathcal{L}_{\text{entro}} \tag{9}$$

where $\mathcal{L}_{\text{task}}$ is the primary task loss (e.g., cross-entropy for medical image classification), and α is a hyperparameter balancing the entropy regularization. All model parameters, including cluster embeddings, membership projections, degree corrections, and transformer weights, are optimized jointly in an end-to-end manner using stochastic gradient descent.

Experiments

Experiments Setup

Dataset. We apply our proposed DCMM-Transformer for the image classification task to evaluate its performance. We consider the following five diverse and representative medical imaging datasets: ChestXray, SIIM-ACR, INbreast, ADNI, and EyeDisease, each addressing different anatomical structures and imaging modalities.

(1) **ChestXray**: The ChestX-ray dataset (Wang et al. 2017) contains large-scale chest X-ray images. Here we focus on images of two major categories (5,856 in total): pneumonia and normal. (2) **SIIM-ACR**: The SIIM-ACR dataset (Stephens 2019) includes 1,250 labeled radiographic images, classified as either normal lung or collapsed lung. (3) **INbreast**: The INbreast database (Moreira et al. 2012) consists of 6,154 mammography images collected at a breast center in Porto, Portugal. Each image is annotated with one of three classification labels: malignant, benign, and normal. (4) **ADNI**: We use structural connectivity (SC) matrices derived from MRI scans in the publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (Yu et al. 2023c). The final dataset includes 282 cognitively normal (CN) subjects and 149 mild cognitive impairment (MCI) subjects. (5) **EyeDisease**: The EyeDisease dataset (Zanlorensi et al. 2022) contains approximately 4,000 high-resolution retinal (fundus) images, aggregated from multiple public sources. Each image is labeled with one of four categories: Normal, Diabetic Retinopathy (DR), Cataract, and Glaucoma, with roughly 1,000 images per class.

Baselines. We compare with six state-of-the-art ViT baselines: ViT (Dosovitskiy et al. 2021), DeiT (Touvron et al. 2021), Swin Transformer (Liu et al. 2021), MaxViT (Tu et al. 2022), ConViT (d'Ascoli et al. 2021), and SBM-Transformer (Cho et al. 2022). These models cover a broad spectrum of design philosophies, including standard transformers (ViT, DeiT), hierarchical and locality-aware attention (Swin, MaxViT, ConViT), and explicit community structure modeling (SBM-Transformer). This selection ensures that our evaluation is comprehensive.

Training Details. Since the dataset sizes are medium to small in scale, we build our DCMM-Transformer on ViT-Small (ViT-S) and initialize it with pre-trained weights (Dosovitskiy et al. 2021). For fair comparison, all baseline models also use the small variant with pretrained weights. We use the AdamW optimizer, with a training batch size of 16 and a total of 100 training epochs. The learning rate is linearly increased from 0 to 0.0005, and subsequently follows a cosine decay schedule. Other training hyperparameters (such as dropout rate) are kept consistent

across all experiments and methods for fair comparison. Our DCMM-Transformer introduces three specific hyperparameters: the entropy regularization weight α, the community bias strength λ, and the number of communities L. After the grid search, we set the hyperparameters to $\alpha = 0.1$, $\lambda = 10$, and $K = 100$. These values are used as defaults for DCMM-Transformer, but all three hyperparameters can be further optimized using cross-validation. All experiments are conducted on NVIDIA H100 GPUs.

Exepriment Results

Table 1 presents classification accuracies on five datasets. The DCMM-Transformer consistently outperforms all baselines across all datasets, achieving an average accuracy of 87.3%, surpassing the next-best method (MaxViT, 85.8%) by a significant margin. Notably, DCMM demonstrates the strongest gains on the ADNI dataset (74.1%), highlighting its capacity to capture subtle structural variations in brain connectivity graphs, a challenging task where standard vision transformers underperform. The model also achieves state-of-the-art performance on the ChestXray (96.0%) and SIIM-ACR (88.2%) datasets, showcasing its effectiveness on 2D radiographic images with diverse pathological characteristics. These results validate the ability of the proposed degree-corrected mixed-membership attention mechanism to enhance representation learning in medical image classification tasks.

It is worth noting that the SBM-Transformer underperforms relative to standard baselines in this setting. This is likely because SBM-Transformer was originally designed to promote sparsity and reduce computational cost in NLP tasks, rather than to model the nuanced spatial and anatomical structures present in medical images. In contrast, the DCMM-Transformer is explicitly designed to exploit the spatial consistency and community-like organization of medical image patches. Moreover, in clinical practice, model performance and interpretability are typically prioritized over computational cost, reinforcing the practical value of our approach. These results validate the effectiveness of the proposed degree-corrected mixed-membership attention mechanism in enhancing representation learning for medical image classification.

Interpretability

Beyond improving classification accuracy, the proposed DCMM-Transformer mechanism significantly improves the interpretability of ViTs on medical imaging tasks. To assess interpretability, we visualize the attention maps generated by our DCMM-Transformer as well as those from standard attention mechanisms across these datasets. Figure 3 illustrates representative results on the five medical datasets. Each row in the figure corresponds to a specific diagnostic category (such as malignant, benign, or normal for INbreast). For each category, two representative image samples are shown. In each sample, the first column displays the original input image, followed by the attention heatmap generated by a standard ViT, and finally the attention map produced by our DCMM-Transformer.

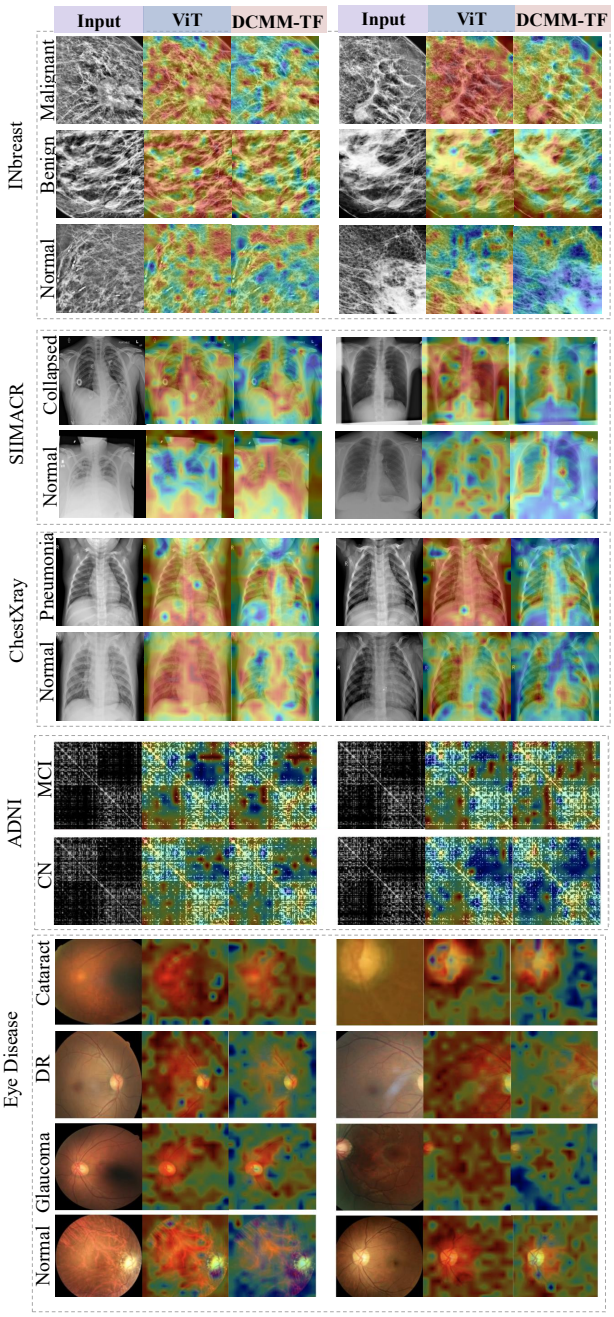

Figure 3: Visualization of attention maps from a standard Transformer and the proposed DCMM-Transformer (abbreviated as DCMM-TF in the figure) on the five datasets. The left column shows the original input images, the middle column displays attention maps from the standard Transformer, and the right column presents attention maps from DCMM-Transformer. For each category, two random subjects are displayed. DR is short for Diabetic Retinopathy.

DCMM-Transformer demonstrates clinically superior interpretability, as evidenced by the following observation exam-

Method	ChestXray	SIIMARC	INbreast	ADNI	EyeDisease	Average
DCMM-Transformer	96.0	88.2	85.1	74.1	93.3	87.3
w/o $\mathcal{L}_{\text{entro}}$	95.2	84.2	83.8	66.7	92.3	84.4
w/o DC	94.9	86.9	83.8	66.7	92.9	85.0
w/o MM	95.6	87.5	81.8	65.4	93.0	84.7

Table 2: Ablation study on the effect of key components in DCMM-Transformer across five medical imaging datasets. w/o $\mathcal{L}_{\text{entro}}$: without the structural constraint loss; w/o DC: without the degree correction module; w/o MM: without the mixed-membership module.

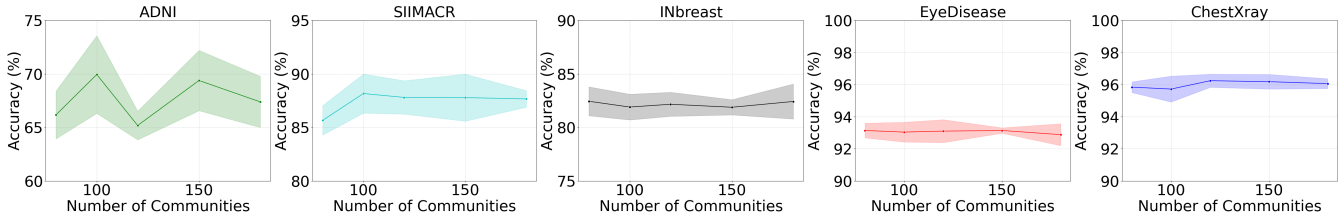

Figure 4: Influence of the number of communities on classification performance across five medical image datasets. The mean classification accuracy and standard deviation over three independent runs for each setting are reported.

ples.

(1) INbreast Mammography: In malignant cases, the DCMM-Transformer attention maps are tightly concentrated around the dense, spiculated masses, which are classic radiological indicators of malignancy, providing precise localization with minimal spillover into adjacent, non-pathological tissue. In contrast, ViT's attention is more diffuse, often extending across wide areas of the breast, including normal fibroglandular tissue and even the pectoral muscle, which reduces clinical clarity. For both benign and normal INbreast cases, DCMM-Transformer continues to produce attention maps that coherently follow the structure of the glandular tissue, showing respect for natural anatomical boundaries.

(2) SIIM-ACR Chest X-ray: For collapsed lung (pneumothorax) cases, the DCMM-Transformer directs attention to the lateral and apical lung zones, precisely where a radiologist would search for the pleural line and absence of peripheral lung markings, key signs of pneumothorax. In particular, the model's attention aligns well with the visible pleural edge and adjacent lucent area, which are crucial for accurate diagnosis. The ViT, by contrast, tends to distribute its attention as a vague cloud across the lung, offering little specific guidance on where to look for pathology. In normal chest X-rays, DCMM-Transformer generally restricts focus within the lung fields and avoids irrelevant areas like the diaphragm and chest wall.

Ablation and Sensitivity Analysis

Ablation. We conducted comprehensive ablation studies to evaluate the impact of the key components of DCMM-Transformer, and the entropy penalty term on model performance across five medical imaging datasets. Each component was removed individually to isolate its effect, and the results are summarized in Table 2. Specifically, omitting the entropy loss leads to the largest average performance drop (from 87.3% to 84.4%), highlighting its critical role in encouraging confident, interpretable community assignments. Eliminating degree correction or mixed-membership also decreases accuracy (to 85.0% and 84.7% average, respectively), underscoring the value of modeling both node connectivity heterogeneity and overlapping community structure.

Sensitivity. Our sensitivity analysis shows that DCMM-Transformer is remarkably stable across a broad range of hyperparameter settings. Among them, the number of communities is particularly important, as it directly and indirectly influences the DCMM bias. We therefore further examine the model's performance under different numbers of communities. As seen in Figure 4, varying L from 80 to 180 yields stable classification accuracy, especially on INbreast, EyeDisease, and ChestXray. This means the model doesn't require precise tuning of this parameter; it works well over a broad range, which is practical in real medical applications.

Conclusion

In this work, we proposed the DCMM-Transformer, a novel Vision Transformer architecture that integrates the DCMM community structure into the attention mechanism. By learning soft group memberships and degree scalars directly from the query representations, and integrating them into the attention mechanism via an additive bias, DCMM-Transformer enables the model to focus on semantically coherent and clinically relevant regions within medical images. This approach not only enhances classification accuracy but also improves interpretability. Extensive experiments across diverse medical imaging datasets demonstrate that DCMM-Transformer consistently outperforms strong transformer baselines, while ablation studies validate the individual contributions of the degree correction, mixed-membership, and entropy regularization modules.

Acknowledgments

This work was partially supported by the U.S. National Science Foundation (NSF) under grants DMS-1903226, DMS-1925066, DMS-2124493, DMS-2311297, DMS-2319279, DMS-2318809, and by the U.S. National Institutes of Health (NIH) under grant R01GM152814. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, or NIH.

References

Airoldi, E. M.; Blei, D.; Fienberg, S.; and Xing, E. 2008. Mixed membership stochastic blockmodels. *Advances in neural information processing systems*, 21.

Chavoshnejad, P.; Chen, L.; Yu, X.; Hou, J.; Filla, N.; Zhu, D.; Liu, T.; Li, G.; Razavi, M. J.; and Wang, X. 2021. An integrated finite element method and machine learning algorithm for brain morphology prediction. *Cerebral Cortex*, 33(15): 9354–9366.

Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. 2024. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97: 103280.

Cheng, H.-M.; Ning, Y.-Z.; Yin, Z.; Yan, C.; Liu, X.; and Zhang, Z.-Y. 2018. Community detection in complex networks using link prediction. *Modern Physics Letters B*, 32(01): 1850004.

Cho, S.; Min, S.; Kim, J.; Lee, M.; Lee, H.; and Hong, S. 2022. Transformers meet stochastic block models: attention with data-adaptive sparsity and cost. *Advances in Neural Information Processing Systems*, 35: 24706–24719.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

d'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; and Sagun, L. 2021. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning*, 2286–2296. PMLR.

Esteva, A.; Chou, K.; Yeung, S.; Naik, N.; Madani, A.; Mottaghi, A.; Liu, Y.; Topol, E.; Dean, J.; and Socher, R. 2021. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1): 5.

Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H. R.; and Xu, D. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 574–584.

Holland, P. W.; Laskey, K. B.; and Leinhardt, S. 1983. Stochastic blockmodels: First steps. *Social networks*, 5(2): 109–137.

Huang, Z.; Yu, X.; Wessler, B. S.; and Hughes, M. C. 2025. Semi-supervised multimodal multi-instance learning for aortic stenosis diagnosis. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 1–5.

Huang, Z.; Yu, X.; Zhu, D.; and Hughes, M. C. 2024. Interlude: Interactions between labeled and unlabeled data to enhance semi-supervised learning. *arXiv preprint arXiv:2403.10658*.

Ibtehaz, N.; Yan, N.; Mortazavi, M.; and Kihara, D. 2024. ACC-ViT: Atrous Convolution's Comeback in Vision Transformers. *arXiv preprint arXiv:2403.04200*.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136. PMLR.

Jin, J.; Ke, Z. T.; and Luo, S. 2024. Mixed membership estimation for social networks. *Journal of Econometrics*, 239(2): 105369.

Karrer, B.; and Newman, M. E. 2011. Stochastic blockmodels and community structure in networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 83(1): 016107.

Li, J.; Chen, J.; Tang, Y.; Wang, C.; Landman, B. A.; and Zhou, S. K. 2023a. Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical image analysis*, 85: 102762.

Li, Z.; Cong, Y.; Chen, X.; Qi, J.; Sun, J.; Yan, T.; Yang, H.; Liu, J.; Lu, E.; Wang, L.; et al. 2023b. Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors. *IScience*, 26(1).

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J. A.; Van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88.

Liu, X.; Cheng, H.-M.; and Zhang, Z.-Y. 2019. Evaluation of community detection methods. *IEEE Transactions on Knowledge and Data Engineering*, 32(9): 1736–1746.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Moreira, I. C.; Amaral, I.; Domingues, I.; Cardoso, A.; Cardoso, M. J.; and Cardoso, J. S. 2012. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2): 236–248.

Stephens, K. 2019. ACR, SIIM Name Winners of Pneumothorax Detection Machine Learning Challenge. *AXIS Imaging News*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jegou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, 10347–10357.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, 459–479. Springer.

Varoquaux, G.; and Cheplygina, V. 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1): 48.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Chen, H.; Gan, C.; Lin, H.; Dou, Q.; Tsougenis, E.; Huang, Q.; Cai, M.; and Heng, P.-A. 2019. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE transactions on cybernetics*, 50(9): 3950–3962.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106.

Yu, X.; Hu, D.; Zhang, L.; Huang, Y.; Wu, Z.; Liu, T.; Wang, L.; Lin, W.; Zhu, D.; and Li, G. 2022a. Longitudinal infant functional connectivity prediction via conditional intensive triplet network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 255–264. Springer.

Yu, X.; Huang, Z.; Xue, Y.; Zhang, L.; Liu, T.; and Zhu, D. 2023a. Noisynn: Exploring the influence of information entropy change in learning systems. *arXiv preprint arXiv:2309.10625*.

Yu, X.; Scheel, N.; Zhang, L.; Zhu, D. C.; Zhang, R.; and Zhu, D. 2021. Free water in T2 FLAIR white matter hyperintensity lesions. *Alzheimer's & Dementia*, 17: e057398.

Yu, X.; Wu, Z.; Zhang, L.; Zhang, J.; Lyu, Y.; and Zhu, D. 2024a. Cp-clip: Core-periphery feature alignment clip for zero-shot medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 88–97. Springer.

Yu, X.; Zhang, J.; Chen, T.; Zhuang, Y.; Chen, M.; Cao, C.; Lyu, Y.; Zhang, L.; Su, L.; Liu, T.; and Zhu, D. 2025. Domain-Adaptive Diagnosis of Lewy Body Disease with Transferability Aware Transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 184–193. Springer.

Yu, X.; Zhang, L.; Dai, H.; Lyu, Y.; Zhao, L.; Wu, Z.; Liu, T.; and Zhu, D. 2023b. Core-periphery principle guided redesign of self-attention in transformers. *arXiv preprint arXiv:2303.15569*.

Yu, X.; Zhang, L.; Lyu, Y.; Liu, T.; and Zhu, D. 2023c. Supervised deep tree in Alzheimer's disease. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Yu, X.; Zhang, L.; Wu, Z.; and Zhu, D. 2024b. Core-periphery multi-modality feature alignment for zero-shot medical image analysis. *IEEE Transactions on Medical Imaging*.

Yu, X.; Zhang, L.; Zhao, L.; Lyu, Y.; Liu, T.; and Zhu, D. 2022b. Disentangling spatial-temporal functional brain networks via twin-transformers. *arXiv preprint arXiv:2204.09225*.

Zanlorensi, L. A.; Laroca, R.; Luz, E.; Britto Jr, A. S.; Oliveira, L. S.; and Menotti, D. 2022. Ocular recognition databases and competitions: A survey. *Artificial Intelligence Review*, 55(1): 129–180.

Zhang, J.; Yu, X.; Lyu, Y.; Zhang, L.; Chen, T.; Cao, C.; Yan, Z.; Chen, M.; Liu, T.; and Zhu, D. 2025. BrainAdapter: Enhancing Neurological Disorder Analysis with Adapter-Tuning Multimodal Large Language Models. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 1–5.