LEARNING LABEL-EFFICIENT INTERPRETABLE MEDICAL IMAGE DIAGNOSIS VIA SEMI-SUPERVISED HYPERGRAPH CONCEPT BOTTLENECK MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning has revolutionized medical image analysis, delivering exceptional diagnostic accuracy across diverse applications. Yet, the lack of interpretability in its decision-making hinders clinical adoption, particularly in high-stakes medical contexts where transparency is paramount for trustworthiness. For example, in Placenta Accreta Spectrum (PAS), subtle cues in ultrasound imaging challenge reliable diagnosis, rendering black-box models untrustworthy for accurate scoring. To address this, Concept Bottleneck Models (CBMs) offer a promising avenue by embedding clinically meaningful intermediate concepts into the diagnosis pipeline, enabling clinicians to scrutinize and refine model outputs. However, conventional CBMs falter in capturing complex inter-concept dependencies and demand costly, expert-driven concept annotations, limiting their scalability. This study introduces a novel semi-supervised CBM framework designed for medical imaging, which leverages dual-level hypergraph learning to model high-order concept dependencies and generate domain-adaptive pseudo-labels. Our approach achieves superior interpretability and performance by integrating a concept-level hypergraph for enhanced reasoning and an image-level hypergraph for robust pseudo-label generation. Experiments on a newly annotated PAS ultrasound dataset and a breast ultrasound public dataset demonstrate the effectiveness of the proposed concept label-efficient interpretable framework. Its universality is further validated on the dermoscopic image dataset SkinCon. The core code is available at the appendix.

1 INTRODUCTION

Deep learning has driven substantial advancements in medical image analysis, achieving state-of-the-art performance across various diagnostic tasks (Chen et al., 2022; Liu et al., 2022; Zhou et al., 2021). Yet despite this progress, its clinical adoption remains limited, primarily due to the lack of interpretability. In high-stakes medical decision-making, models must not only be accurate, but also provide transparent reasoning that clinicians can understand, validate, and act upon (Tjoa & Guan, 2020; Reddy, 2022; Reyes et al., 2020; Nasarian et al., 2024). For example, *Placenta Accreta Spectrum (PAS)*, a life-threatening pregnancy complication, demands early and accurate diagnosis due to severe risks like hemorrhage. This makes interpretability essential for reliable clinical decision-making. Ultrasound imaging is widely used for PAS considering its non-invasiveness, real-time capability, and cost-effectiveness (Jauniaux et al., 2018; Cali et al., 2019). However, it also presents operator dependency, subtle imaging dynamics, and complex anatomical structures (Sarris et al., 2012; Avola et al., 2021), challenging existing deep models. These facts encourage deep models not only to achieve great results, but also to offer interpretable insights that align with clinical reasoning, enabling trustworthy decisions and intervention in real-world applications.

Concept Bottleneck Models (CBMs)(Koh et al., 2020; Yuksekgonul et al., 2022) provide a promising solution by introducing human-understandable intermediate concepts, allowing clinicians to trace decision pathways and correct mispredictions (Kim et al., 2023b; Pang et al., 2024; Chowdhury et al., 2024). However, existing CBMs face two bottlenecks: **(1) Traditional CBMs assume independence among concepts, overlooking essential inter-concept relationships that are instead inherent in medical imaging and critical for holistic reasoning.** For example, lesion morphol-

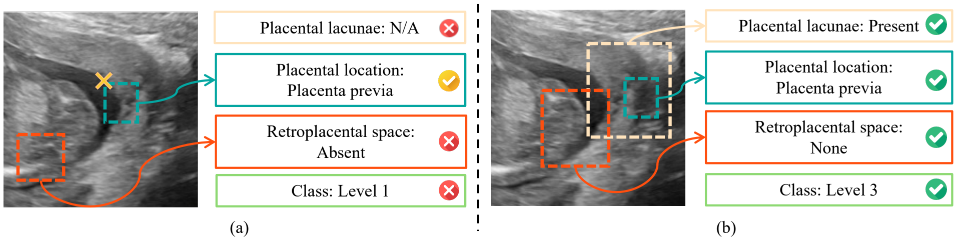
Placental lacunae: N/A
Placental location: Placenta previa
Retroplacental space: Absent
Class: Level 1
(a)
Placental lacunae: Present
Placental location: Placenta previa
Retroplacental space: None
Class: Level 3
(b)

Figure 1: **Traditional methods degenerate in a semi-supervised spirit.** The conventional CEM (a) and our HyperCBM (b) try to infer the PAS severity level from the predicted concepts. CEM illustrates three error modes: ignoring lacunae, misinterpreting the retroplacental space, and focusing on a biased placental location. These concept errors yield the wrong severity. Instead, HyperCBM successfully predicts severity from the correct concepts and attentions that match expert assessment.

ogy interpretation depends on surrounding textures, vascular patterns, and anatomical structures. **(2) CBMs typically require resource-intensive and time-consuming concept-level annotations to achieve satisfactory interpretability and maintain decent diagnosis results**. Expensive annotation costs unfortunately impede the scalability and application of CBMs in clinical scenarios. To ease concept annotation, Semi-Supervised Concept Bottleneck Model (SSCBM) has been recently proposed (Hu et al., 2024), leveraging unlabeled data for training. While promising, SSCBM relies on pseudo-labels derived from ImageNet pre-trained features, a strategy that introduces a significant domain gap when applied to clinical imaging. As a result, pseudo-labels lack medical fidelity, undermining both interpretability and downstream diagnostic accuracy (Liu et al., 2023; Li et al., 2020a). ***Addressing this requires a new framework that not only models the semantic dependencies between concepts, but also generates pseudo-labels grounded in domain-specific image semantics, which we tackle in this work.***

In this paper, we propose a novel Hypergraph-driven semi-supervised concept bottleneck framework, dubbed ***HyperCBM***, tailored for label-efficient, interpretable medical image diagnosis. As shown in Fig. 1, HyperCBM is significantly superior to traditional methods like CEM (Espinosa Zarlenga et al., 2022) in detecting clinical concepts and diagnosing under limited concept labels. Our framework tackles the dual challenges of high-order inter-concept relationship modeling and domain-specific pseudo-label generation, thereby enhancing both the interpretability and performance of concept bottleneck models.

Our contributions are summarized as:

- A semi-supervised concept bottleneck framework, which is the first design for medical imaging, improving both label efficiency and interpretability.
- A hypergraph-enhanced concept representation learning (HECRL) introduced to model high-order inter-concept relationships, enhancing diagnostic reasoning accuracy.
- A Hypergraph Image Dynamic Pseudo-labeling (HIDP) generation strategy developed, leveraging adaptive features to robustly exploit unlabeled data.
- Extensive experiments are conducted on a newly curated Placenta Accreta Spectrum dataset, public breast ultrasound and dermoscopic image datasets, demonstrating its effectiveness and state-of-the-art performance against existing methods.

2 Related Work

Concept Bottleneck Models (CBMs) have emerged as a promising approach for enhancing interpretability in deep learning. CBMs introduce an intermediate concept layer between input and prediction, enabling models to make decisions based on human-interpretable concepts. Early works (Koh et al., 2020; Espinosa Zarlenga et al., 2022) demonstrated CBMs could improve generalization and transparency, though they suffered from performance degradation compared to traditional black-box models and required expensive manual annotations. To address this, recent efforts (Chauhan et al., 2023; Oikarinen et al., 2023) proposed interactive CBMs that selectively annotate concepts and label-free CBMs that eliminate the need for labeled concept data. However, they heavily rely on large language models like GPTs, which have reliability issues (Lai et al., 2023), and severely undermine their interpretability. Studies like (Magister et al., 2021; Barbiero et al.,

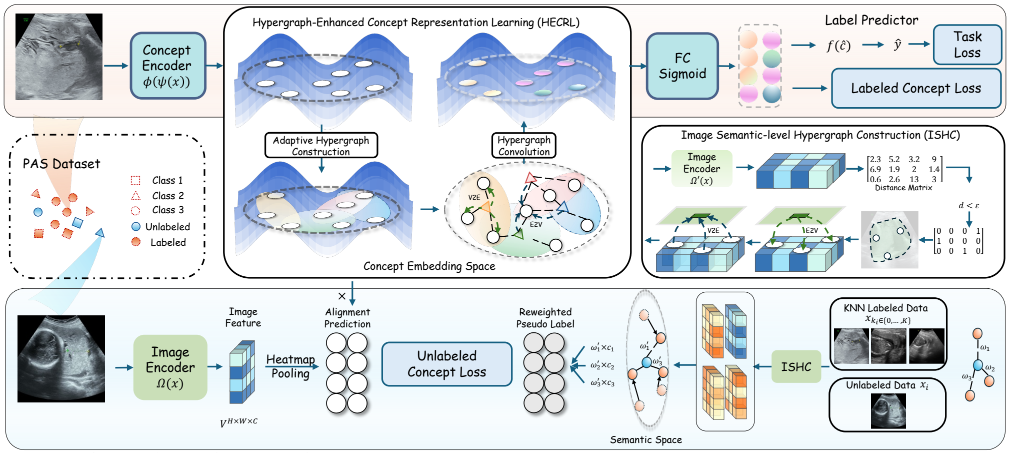

Figure 2: **Overview of HyperCBM**, a hypergraph-driven semi-supervised concept bottleneck model for ultrasound imaging. The framework integrates Hypergraph-Enhanced Concept Representation Learning (HECRL) for high-order inter-concept modeling via adaptive hypergraph propagation, and Hypergraph Image Dynamic Pseudo-labeling (HIDP) for reliable pseudo-label generation.

2024) map graphs to concept spaces using clustering and human-in-the-loop strategies to improve transparency. CBMs work in the image field also include the works of (Havasi et al., 2022; Kim et al., 2023a; Sheth & Ebrahimi Kahou, 2023). Despite this progress, few approaches explore label efficiency, achieving scalable CBMs by semi-supervised learning in clinical applications.

Semi-supervised learning (SSL) is widely used in medical image diagnosis where labeled data is scarce and expensive to obtain. Classical SSL strategies include pseudo-labeling (Kamraoui et al., 2021; Li et al., 2020b), consistency regularization (Gu et al., 2025; Xiao et al., 2025), and hybrid methods like MixMatch (Berthelot et al., 2019) and FixMatch (Sohn et al., 2020). In diagnosis-oriented tasks, POPCORN (Kamraoui et al., 2021) introduces progressive pseudo-labeling guided by feature similarity to improve classification in liver and lung cancer. MemSAM (Deng et al., 2024) utilizes a memory bank of anatomical priors to generate pseudo-masks without manual labels. Graph-based SSL has also shown promise in classification settings. For instance, GraphX-NET (Aviles-Rivero et al., 2019) models chest X-ray data with graph structures, and NoTeacher (Unnikrishnan et al., 2021) combines consistency regularization with probabilistic graphical models to impose structural constraints. While these methods have achieved strong performance in medical diagnosis, *they are predominantly based on black-box models and neglect model transparency and interpretability.*

Visual Graph Learning combines visual representation with structured relational reasoning and has become an effective paradigm for modeling high-order semantics in vision tasks. Unlike traditional CNNs or self-attention models that operate over local pixel or patch-level features, graph-based methods construct relational graphs where each node represents a data instance or image region, and edges encode semantic affinities (Zhu, 2005; Chong et al., 2020; Song et al., 2022). These approaches allow models to incorporate both explicit and latent relationships, offering a more structured understanding of similarity. Recent works incorporate hypergraphs to capture richer multi-way relationships beyond pairwise interactions. Hypergraph-based methods (Gao et al., 2012; Huang et al., 2009) have been applied to GNNs (Han et al., 2023; Srinivas et al., 2024) for better representation learning. HgVT (Fixelle, 2025) propose dynamic hypergraph construction guided by feature similarity and regularization strategies, enabling structure-aware learning directly within vision transformer architectures. These advances demonstrate the potential of structured learning for enhanced generalization in complex visual recognition tasks. However, none of them explore inter-concept relationships for interpretability.

3 Method

In this section, we introduce our method **HyperCBM** enabling explainable and data-efficient medical image diagnosis. We first present its preliminaries. Then, we provide a detailed description of the **Hypergraph-Enhanced Concept Representation Learning** (HECRL) and **Hypergraph Image Dynamic Pseudo-labeling** (HIDP) components. The overall framework is illustrated in Fig. 2.

3.1 Preliminary: CBM and SSCBM

CBM (Koh et al., 2020) is a class of interpretable models designed to predict a target $y \in Y$ from an input $x \in X$ through an intermediate concept space C. The concept set $C = \{p_1, \dots, p_t\}$ consists of t binary concepts provided by experts. The training dataset is represented as $D = \{(x(i), y(i), c(i))\}_{i=1}^N$, where for each sample i, $x(i) \in X \subseteq \mathbb{R}^d$ is the input image, $y(i) \in Y \subseteq \mathbb{R}^l$ denotes the label with l classes, and $c^{(i)} = (c_i^1, \cdots, c_i^t) \in \{0,1\}^t$ is the concept vector indicating the presence or absence of each concept. CBM learns two mappings: a **concept encoder** $g : \mathbb{R}^d \to \mathbb{R}^t$, which transforms the input x into the concept space $\hat{c} = g(x)$, and a **label predictor** $f : \mathbb{R}^t \to \mathbb{R}^l$, which maps the concept vector $\hat{c}$ to the final prediction $\hat{y} = f(\hat{c})$.

CEM (Espinosa Zarlenga et al., 2022) mitigates CBM-induced performance degradation using high-dimensional concept embeddings. For each input x, CEM generates t concept embeddings $\hat{c}_1, \hat{c}_2, \dots, \hat{c}_t,$ where each concept $\hat{c}_i$ is represented by two embeddings $\hat{c}_i^+, \hat{c}_i^- \in \mathbb{R}^m$, corresponding to the "TRUE" and "FALSE" states of the concept respectively. These embeddings are generated using a DNN $\psi(x)$ to produce a latent representation $h \in \mathbb{R}^n$, followed by concept-specific fully connected layers: $\hat{c}_i = \phi_i(h)$. A differential scoring function $s : \mathbb{R}^{2m} \to [0,1]$ is used to align the embeddings with ground truth concepts, predicting the probability $\hat{p}_i = \sigma(W_s[\hat{c}_i^+, \hat{c}_i^-]^\top + b_s)$ of concept c_i being active. The final concept embedding is computed as:

$$\hat{\boldsymbol{c}}_i := \hat{p}_i \hat{c}_i^+ + (1 - \hat{p}_i)\hat{c}_i^-$$

CEM generates high-quality concept embeddings enriched with semantic information, enhancing both interpretability and task accuracy. However, this benefit comes at the cost of dense expert annotations on concepts. SSCBM (Hu et al., 2024) extends CEM by leveraging limited labeled data with abundant unlabeled data, reducing the concept annotation burden. In SSCBM, the input set $\mathcal{X}$ is divided into two disjoint subsets for semi-supervised setting: $\mathcal{X} = \mathcal{X}_L \cup \mathcal{X}_U$, where $\mathcal{X}_L$ represents a small subset of labeled data and $\mathcal{X}_U$ denotes the remaining unlabeled data, with $|\mathcal{X}_L| \ll |\mathcal{X}_U|$. For $x^{(j)} \in \mathcal{X}_L$, both concept annotations $c^{(j)}$ and class labels $y^{(j)}$ are available. For $x^{(i)} \in \mathcal{X}_U$, only the class label $y^{(i)}$ is accessible. Under these settings, given the combined training dataset $\mathcal{D} = \mathcal{D}_L \cup \mathcal{D}_U,$ where $\mathcal{D}_L = \{(x^{(j)}, y^{(j)}, c^{(j)})\}_{j=1}^{|\mathcal{X}_L|}$ and $\mathcal{D}_U = \{(x^{(i)}, y^{(i)})\}_{i=1}^{|\mathcal{X}_U|},$ the objective is to jointly leverage both labeled and unlabeled data to train a CEM.

3.2 Hypergraph-Enhanced Concept Representation Learning

Traditional CBMs overlook high-order semantic relations between concepts, which are crucial for capturing complex patterns in medical imaging, while the hypergraph structure effectively models multi-way correlations, representing their dependencies beyond the pairwise connections of traditional graphs. Specifically, we construct an image-level concept embedding hypergraph $\mathcal{H}_c = (\mathcal{V}_c, \mathcal{E}_c, \mathcal{W}_c)$ for each image in a batch, where $\mathcal{V}_c$ denotes t concept embeddings in total $\{\hat{\mathbf{c}}^i\}_{i=1}^t$, with each $\hat{\mathbf{c}}^i \in \mathbb{R}^m$, in which m is the dimensionality of a single concept embedding. $\mathcal{H}_c$ is instantiated once per image, while the subsequent HGNN+ layers share weights across all images in a batch. Here, $\mathcal{E}_c$ represents adaptively formed concept clusters, while $\mathcal{W}_c$ reflects the relative importance of each hyperedge. The detailed process is described below.

3.2.1 Adaptive Hypergraph Construction

Considering high-order semantic relationships, we first quantify pairwise relationships between concepts using cosine similarity in each image. This measure ensures that hyperedges are formed among semantically related concepts: $S_{ij} = \frac{\hat{\mathbf{c}}^i \cdot \hat{\mathbf{c}}^j}{\|\hat{\mathbf{c}}^i\| \|\hat{\mathbf{c}}^j\|}$. The resulting similarity matrix $S \in \mathbb{R}^{t \times t}$ serves as the foundation for adaptive hyperedge construction. Traditional methods often rely on fixed k-nearest neighbors, which may not reflect the varying semantic density among concepts. To address this, we define an adaptive target neighborhood size:

$$k_{\text{init}} = \max\left(\lfloor t \times \text{initial\_ratio} \rfloor, k_{\text{min}}\right), \tag{1}$$

where initial_ratio and k_{min} are predefined hyperparameters controlling the neighborhood scale. The outer $\max(\cdot)$ selects the densest local region as an upper bound so that even concepts located in tight clusters retain sufficient neighbours.

To ensure consistent semantic cohesion, we compute a global similarity threshold τ based on the average similarity across the top-k_{init} neighbors:

$$\tau = \frac{1}{t} \sum_{i=1}^{t} \frac{1}{k_{\text{init}}} \sum_{j \in \text{top-}k_{\text{init}}(S_{ij})} S_{ij}. \tag{2}$$

Based on τ, we update the adaptive neighborhood size $\hat{k}$ and construct hyperedges for each concept:

$$\hat{k} = \max\left(\min\left(\max_{i} \sum_{j} \mathbb{1}_{S_{ij} > \tau}, t-1\right), k_{\min}\right), \tag{3}$$

$$e_c^i = \{c^j \mid S_{ij} > \tau \text{ and among top-}\hat{k} \text{ similarities}\}. \tag{4}$$

Each such set $e_i^c \subset \mathcal{V}_c$ forms a hyperedge centered on concept c^i, capturing its most semantically related neighbors under a similarity threshold. Collectively, these egocentric hyperedges define the full hyperedge set $\mathcal{E}_c = \{e_c^1, e_c^2, \ldots, e_c^t\}$, which constitutes the concept-level hypergraph $\mathcal{H}_c = (\mathcal{V}_c, \mathcal{E}_c, \mathcal{W}_c)$. This adaptive process ensures that hyperedges capture semantically coherent concept clusters, providing a flexible and robust representation of high-order dependencies.

3.2.2 Attention-driven Hyperedge Weighting

To ensure that the model prioritizes clinically relevant concept clusters, we introduce an *attention-driven weighting mechanism*. This mechanism dynamically assigns weight scores to hyperedges, highlighting semantically salient relationships.

For each hyperedge $e \in \mathcal{E}_c$ that connects $\hat{k}$ concepts, we aggregate the corresponding concept embeddings $C_c^e \in \mathbb{R}^{\hat{k} \times m}$ and project the embeddings into a shared d_a-dimensional latent space:

$$Q = W_Q C_c^e, K = W_K C_c^e, V = W_V C_c^e, \tag{5}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{m \times d_a}$ are learnable parameters and $Q, K, V \in \mathbb{R}^{\hat{k} \times d_a}$. We compute attention scores α_c^e to capture the importance of each hyperedge by its internal semantic consistency: $\alpha_c^e = \text{softmax}\left(QK^\top / \sqrt{d}\right) \in \mathbb{R}^{\hat{k} \times \hat{k}}$. α_c^e reflects the alignment of concept embeddings within a hyperedge. We then derive a scalar, unnormalised importance by aggregating the attention-reweighted value features across nodes within the hyperedge. Specifically, we compute the mean of attention-weighted value features $V_c^e \in \mathbb{R}^{\hat{k} \times d_a}$ to obtain a representative feature vector $\bar{V}_c^e \in \mathbb{R}^{d_a}$ that captures the collective semantic information.

$$\bar{V}_c^e \;=\; \frac{1}{\hat{k}} \sum_{i=1}^{\hat{k}} \bigl(\alpha_c^e V\bigr)_{i,:}, \quad w_c^e \;=\; \bigl\|\bar{V}_c^e\bigr\|_2 \tag{6}$$

The ℓ_2-norm then quantifies the semantic coherence and feature activation intensity of this aggregated representation, serving as a measure of the hyperedge's clinical relevance and conceptual consistency. We then apply a softmax operation over all hyperedges belonging to the same image. This attention-driven weighting mechanism ensures that the model dynamically focuses on concept clusters that are more relevant to clinical outcomes:

$$\tilde{w}_c^e \;=\; \frac{\exp\bigl(w_c^e\bigr)}{\sum_{e' \in \mathcal{E}_c} \exp\bigl(w_c^{e'}\bigr)}. \tag{7}$$

The normalised coefficients $\{\tilde{w}_c^e\}_{e \in \mathcal{E}_c}$ populate the diagonal matrix $\mathcal{W}_c$, which is used as the hyperedge weight matrix in subsequent HGNN$^+$ propagation.

After constructing the hypergraph with adaptively weighted hyperedges, we employ *HGNN+ layers* (Gao et al., 2022) for high-order semantic reasoning. Unlike traditional HGNN (Feng et al., 2019), HGNN+ integrates vertex-to-hyperedge (V2E) aggregation and hyperedge-to-vertex (E2V) propagation into a single compact formulation. Given the concept embeddings $\hat{\boldsymbol{c}}_{\boldsymbol{i}}^{(l)}$ at the l-th layer, the HGNNConv+ layer is defined as:

$$\hat{\boldsymbol{c}}_{\boldsymbol{i}}^{(l+1)} = \sigma\left(D_v^{-1} H \mathcal{W}_c D_e^{-1} H^\top \hat{\boldsymbol{c}}_{\boldsymbol{i}}^{(l)} \Theta^{(l)}\right) + \hat{\boldsymbol{c}}_{\boldsymbol{i}}^{(l)}, \tag{8}$$

where H is the hypergraph incidence matrix encoding V2E relations, D_v and D_e are the vertex and hyperedge degree matrices for E2V propagation, $\Theta^{(l)}$ represents the learnable parameters at layer l, and $\sigma(\cdot)$ is a non-linear activation function. The final layer result can be denoted as $\hat{\boldsymbol{c}}_{\text{concept}}$. Note that residual connections are applied to enhance training stability and mitigate over-smoothing. The resulting $\hat{\boldsymbol{c}}_{\text{concept}}$ is forwarded to unlabelled data concept prediction and subsequently to the final classification head.

3.3 Hypergraph Image Dynamic Pseudo-labeling

To overcome the limitations of pseudo-labeling in SSCBM (Hu et al., 2024), particularly the lack of domain-specific semantics, we propose Hypergraph Image Dynamic Pseudo-labeling (HIDP). HIDP constructs an image-level hypergraph on semantic feature maps to improve pseudo-label selection through hypergraph-based feature aggregation, dynamic pseudo-label generation, and enhanced alignment.

3.3.1 Hypergraph Construction on Semantic Feature Maps

Given an unlabeled image sample x, we extract its semantic image feature map as a vertex using an image encoder. To incorporate contextual relationships, we select the K-nearest labeled samples based on similarity in the feature space. To capture high-order spatial correlations among these feature maps, we construct an image-level hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$. Following a distance-based hypergraph paradigm (Feng et al., 2024), each spatial feature point (p, q) in the $h \times w$ grid corresponds to a vertex in the hypergraph. For each vertex v, an ϵ-ball hyperedge includes all neighboring vertices u satisfying a distance constraint in the feature space. This construction ensures that each hyperedge represents a local semantic neighborhood within the image feature space. Note that hyperedges are confined to a single image; cross-image information is introduced only through the subsequent pseudo-label aggregation step.

3.3.2 Hypergraph-driven Dynamic Pseudo-label Generation

To model high-order dependencies among feature points, we apply a single hypergraph convolutional layer with residual connections as:

$$\hat{V} = V + D_v^{-1} H \mathcal{W} D_e^{-1} H^{\top} V \Theta, \tag{9}$$

where V represents vertex features, H the hypergraph incidence matrix, D_v and D_e the vertex and hyperedge degree matrices, respectively, and Θ the learnable parameters. This operation enables each vertex to aggregate information from its connected hyperedges, capturing contextual semantic relationships within the image. Then we compute the Euclidean distance $d_i = \|\hat{V}^u - \hat{V}_i^l\|$ between the feature of the unlabeled sample $\hat{V}^u$ and its K neighbors $\{\hat{V}_i^l\}_{i=1}^K$ with labeled concepts $\{c_i^l\}_{i=1}^K$. The final pseudo-label $\hat{c}_{\text{pseudo}}$ is calculated as:

$$\omega_i = \frac{(1/d_i)}{\sum_{j=1}^{K} (1/d_j)}, \hat{\boldsymbol{c}}_{\text{pseudo}} = \sum_{i=1}^{K} \omega_i \cdot c_i^l. \tag{10}$$

3.3.3 Semi-supervised Training

To enhance concept interpretability while maintaining classification performance, for labeled data, we define concept loss $\mathcal{L}_c$ using binary cross-entropy (BCE) to enforce consistency between predicted concepts $\hat{\boldsymbol{c}}$ and ground-truth labels $\boldsymbol{c}$. For unlabeled data, we introduce the alignment loss $\mathcal{L}_{align}$ to enforce coherence between concept embeddings and image features. We derive $\hat{\boldsymbol{c}}_{\text{hyper}}$ from the semantic feature map V and updated concept embeddings $\hat{\boldsymbol{c}}_{\text{concept}}$ via heatmap-based operations and align it with similarity-based concept labels $\hat{\boldsymbol{c}}_{\text{pseudo}}$ using BCE as $\mathcal{L}_{align}$. The task loss $\mathcal{L}_{task}$ ensures accurate classification by mapping concept embeddings $\hat{\boldsymbol{c}}_{\text{concept}}$ to final predictions $\hat{\boldsymbol{y}}$ via a label predictor $g(\cdot)$, optimized with categorical cross-entropy. The overall objective is formulated as

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_{align}, \tag{11}$$

where λ_1 and λ_2 balance interpretability and classification accuracy in semi-supervised setting (*See Implementation Details*). We define $\mathcal{L}_c = \text{BCE}(\hat{c}, c)$ for labeled samples and $\mathcal{L}_{\text{align}} = \text{BCE}(\hat{c}_{\text{concept}}, \hat{c}_{\text{pseudo}})$ for unlabeled ones, where $\hat{c}_{\text{concept}}$ is obtained via attention-based decoding.

Table 1: Results of concept and task accuracy and AUC across PAS, BrEaST, and SkinCon datasets with different labeled data ratios. '*' denotes the model trained in a fully supervised setting.

Method	Labeled Ratio	*PAS*				*BrEaST*				*SkinCon*			
		Concept Acc.	Class Acc.	Concept AUC	Class AUC	Concept Acc.	Class Acc.	Concept AUC	Class AUC	Concept Acc.	Class Acc.	Concept AUC	Class AUC
HyperCBM	0.01	70.21±1.44	57.33±9.88	52.67±1.97	76.21±7.48	64.26±6.43	65.88±4.57	53.75±2.06	67.32±14.47	86.08±1.06	72.33±1.11	50.52±0.68	57.57±2.54
	0.1	81.61±0.88	76.89±3.85	64.43±1.27	88.40±2.51	72.49±3.90	65.49±7.08	56.66±3.96	70.31±9.70	88.34±0.51	73.99±0.96	54.59±2.03	70.53±3.09
	0.2	83.07±0.70	76.59±2.84	67.08±2.05	90.57±1.60	71.99±2.49	66.27±9.06	55.26±3.51	72.00±8.31	89.06±0.25	75.76±0.99	56.71±1.54	74.39±1.68
	0.4	84.19±1.11	78.82±4.18	68.20±3.00	90.48±3.03	76.25±2.81	75.29±4.22	61.93±4.69	80.11±2.42	89.87±0.22	75.86±0.80	60.14±1.23	75.34±0.98
	0.6	85.57±0.83	80.15±1.65	71.25±1.91	93.33±2.46	77.81±1.40	73.33±7.40	62.70±4.70	80.75±5.15	90.13±0.45	75.48±0.29	61.03±0.89	77.64±1.77
	0.8	85.57±0.86	81.93±3.49	71.89±1.36	91.95±1.44	76.08±1.83	64.70±9.92	57.52±4.85	74.49±7.60	90.12±0.34	75.92±0.93	60.73±1.86	75.30±1.14
HyperCBM*	1.0	**84.93±0.93**	**79.85±2.36**	**71.45±1.90**	**92.51±2.18**	**78.88±1.33**	**74.12±7.78**	**64.13±3.34**	80.61±6.82	**90.43±0.22**	77.48±1.20	**62.34±0.61**	**79.86±1.45**
CBM*	1.0	79.85±2.50	75.85±10.51	64.64±7.29	88.91±8.34	76.30±0.29	70.98±10.03	59.08±2.55	79.88±4.65	90.24±0.32	72.43±1.21	61.72±0.83	69.37±1.73
CEM*	1.0	78.14±2.24	77.78±2.52	61.67±2.66	91.09±1.52	76.36±3.45	71.76±6.40	58.92±5.81	**81.45±2.58**	90.33±0.33	76.10±2.02	62.22±0.82	78.42±1.33
ResNet*	–	N/A	78.22±2.55	N/A	91.09±2.09	N/A	74.12±3.37	N/A	80.50±3.85	N/A	**77.88±0.92**	N/A	77.55±0.80

Table 2: Evaluation with concept labeled ratios of 0.1, 0.4 on three datasets. Best results are in **bold**, second-best are underlined.

Labeled Ratio	Method	*PAS*				*BrEaST*				*SkinCon*			
		Concept Acc.	Class Acc.	Concept AUC	Class AUC	Concept Acc.	Class Acc.	Concept AUC	Class AUC	Concept Acc.	Class Acc.	Concept AUC	Class AUC
0.1	CBM	77.54±1.77	65.93±11.52	57.63±5.05	83.14±9.55	68.16±0.96	61.96±5.63	52.67±1.97	47.02±12.94	88.17±0.44	72.43±1.21	50.48±0.11	53.70±3.68
	CEM	71.23±1.82	69.92±3.88	53.52±2.82	86.03±3.06	68.13±2.28	61.57±5.35	50.79±2.51	58.08±7.19	87.94±0.94	**74.14±0.14**	53.54±2.81	**70.69±3.54**
	SSCBM	73.77±2.55	73.78±2.99	56.28±4.04	**88.68±1.85**	71.43±3.09	61.57±5.77	52.90±4.26	62.87±10.62	87.97±0.99	73.92±0.63	53.08±2.47	69.93±3.59
	Ours	**81.61±0.88**	**76.89±3.85**	**64.43±1.27**	88.40±2.51	**72.49±3.90**	**65.49±7.08**	**56.66±3.96**	**70.31±9.70**	**88.34±0.51**	73.99±0.96	**54.59±2.03**	70.53±3.09
0.4	CBM	79.33±2.14	73.63±9.14	62.72±6.01	87.66±8.81	70.93±1.07	61.57±5.35	52.28±2.01	50.86±7.83	89.79±0.18	72.43±1.21	54.93±1.77	64.25±4.13
	CEM	70.13±1.60	70.37±3.71	51.47±1.94	85.16±3.53	73.50±1.78	66.28±8.54	54.97±2.68	80.45±4.49	88.82±1.22	74.49±1.11	57.42±3.19	72.69±3.66
	SSCBM	78.16±3.81	76.44±1.65	60.51±3.67	89.96±2.02	73.89±2.06	72.94±6.83	56.17±3.23	**81.46±2.18**	88.78±1.31	74.89±0.76	56.46±2.83	72.16±3.63
	Ours	**84.11±1.11**	**78.67±4.18**	**68.73±3.00**	**92.41±5.91**	**76.25±2.81**	**75.29±4.22**	**61.93±4.69**	80.11±2.42	**89.87±0.22**	**75.86±0.80**	**60.14±1.23**	**75.34±0.98**

4 EXPERIMENTS

4.1 DATASETS AND SETTINGS

We evaluate our framework on three datasets: a newly collected ultrasound Placenta Accreta Spectrum grading dataset (***PAS***), a public ultrasound breast lesion diagnosis dataset ***BrEaST*** (Pawłowska et al., 2024), and a public dermoscopic imaging dataset ***Skincon*** (Daneshjou et al., 2022).

PAS is a newly collected placenta-accreta-spectrum ultrasound dataset that contains 671 scans acquired from multiple vendors and annotated with three severity levels and 45 clinically curated concepts. Candidate concepts are extracted by HuatuoGPT-Vision and confirmed by two senior obstetric radiologists. *The detailed dataset statistics are described in Appendix.* **BrEaST** originally contains 256 breast ultrasound scans with 7 concepts from BI-RADS descriptors and 3 diagnostic labels. Following (Wang et al., 2024), we exclude the Normal category, and finally use the 254 abnormal images with malignant and benign categories. **SkinCon** selects 3,205 dermoscopic images from Fitzpatrick-17k (Groh et al., 2021), covering Malignant, Benign, and Non-neoplastic lesions. Two dermatologists densely labelled 48 clinical concepts, of which the 22 occurring in at least 50 images are retained, consistent with (Wang et al., 2024). For all three datasets, images are centre-cropped and resized to 224×224; data are randomly split into training, validation, and test subsets in a 7 : 1 : 2 ratio. Every experiment is repeated with five fixed random seeds, and results are reported as mean ± standard deviation.

We compare with the following baselines: CBM (Koh et al., 2020), CEM (Espinosa Zarlenga et al., 2022), and SSCBM (Hu et al., 2024) in full-supervised and semi-supervised settings. For evaluation, we use the Area Under the Receiver Operating Characteristic Curve (AUC), Accuracy (ACC) as disease diagnosis and concept detection tasks.

Implementation Details. All experiments were conducted on a single NVIDIA RTX 4090 GPU. We employ a ResNet (He et al., 2016) backbone as the image encoder, specifically using ResNet34 for the *BrEaST* and *PAS* datasets, and ResNet50 for the *SkinCon* dataset. Both the concept adapters and the aggregator are implemented as fully-connected layers. The image encoder within the concept encoder, denoted as $\psi(x)$, and the unlabeled image encoder $\Omega(x)$ share their weights. To strictly enforce the stability of pseudo-label generation during training, the image encoder $\Omega^{'}(x)$ utilized to extract features for ISHC module remains frozen throughout each training epoch. Then the weights are updated at the end of each training epoch by synchronizing the latest optimized parameters of $\psi(x)$. We train all models end-to-end utilizing the Adam optimizer. The models for *BrEaST*, *PAS*, and *SkinCon* are trained for a maximum of 150, 250, and 100 epochs, respectively. The corresponding learning rates are set to 5×10^{-4} for *BrEaST* and *PAS*, and 5×10^{-3} for *SkinCon*. To prevent overfitting, we employ an early stopping strategy with a patience of 5 epochs, monitoring the validation loss. For HECRL, the number of nearest neighbors for a hyperedge, $k_{\min}$, is set to 2 for *BrEaST* and 3 for the more concept-diverse *PAS* and *SkinCon* datasets. The radius for the ϵ-ball hyperedge,

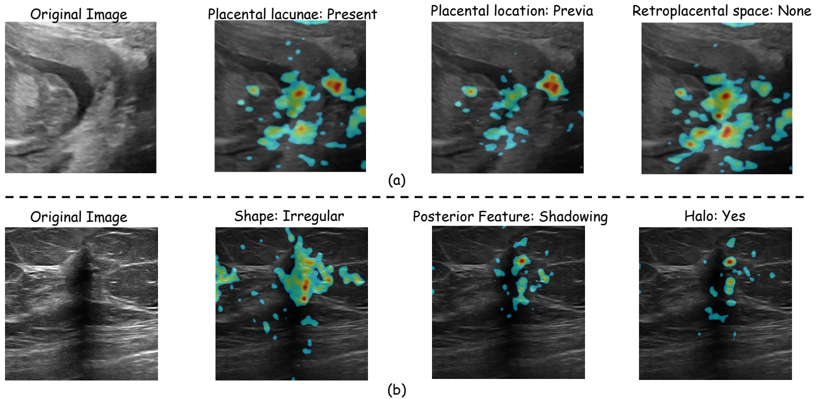
Original Image
Placental lacunae: Present
Placental location: Previa
Retroplacental space: None
(a)
Original Image
Shape: Irregular
Posterior Feature: Shadowing
Halo: Yes
(b)

Figure 3: **Interpretability Visualization**: (a) Concept saliency maps on the PAS dataset, highlighting learned concepts (*e.g.*, placental lacunae, retroplacental space). (b) Concept saliency maps on the BrEaST dataset, capturing key diagnostic features (*e.g.*, irregular shape, posterior features).

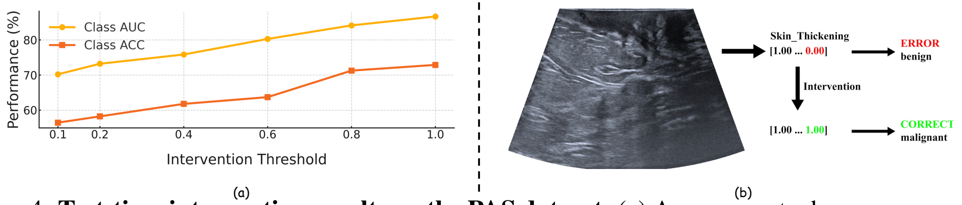

Figure 4: **Test-time intervention results on the PAS dataset**: (a) Any concept whose score exceeds the intervention threshold is forced to zero. This intervention causes a nearly monotonic degradation in diagnosis. (b) An example demonstration of test-time intervention, where correcting "Skin Thickening" shifts the prediction from benign to malignant, improving diagnosis results and demonstrating model applicability.

ϵ is fixed at 6 across all datasets. The loss balancing weights (λ_1,λ_2) are empirically configured to (0.5,0.1), (1.0,0.1), and (2.0,1.0) for the *BrEaST*, *PAS*, and *SkinCon* datasets, respectively. (***See Appendix for details***)

4.2 Quantitative Comparison

In Tab. 1, we present a comprehensive evaluation of HyperCBM's performance in a semi-supervised setting, systematically varying the proportion of labeled data from 1% to 80% across three distinct datasets: PAS, BrEaST, and SkinCon. The results are benchmarked against several fully supervised models, including our own HyperCBM, traditional CBM, CEM, and a standard ResNet backbone. On the PAS dataset, HyperCBM shows a clear and consistent improvement with increased supervision concept label ratio. Concept accuracy rises from 70.21% to 85.57%, and class accuracy increases from 57.33% to 81.93%. A key finding is the model's high data efficiency: with only 40% of labeled data, HyperCBM's class accuracy (78.82%) is already competitive with the fully supervised ResNet34 model (78.22%) and superior to CBM (75.85%) and CEM (77.78%). The model's performance on the BrEaST and SkinCon datasets further validates its capabilities. On BrEaST, despite some performance fluctuations attributable to the dataset's limited number of concept annotations (7 concepts), HyperCBM with 40% labeled data still surpasses other fully supervised concept models in task accuracy. On SkinCon, the performance progression is more stable, with our model at 20% supervision (75.76% class accuracy) nearing the performance of the fully supervised CEM (76.10%). Collectively, these experiments demonstrate that our semi-supervised HyperCBM is highly effective in low-ratio concept label regimes, achieving performance comparable or superior to fully supervised baselines with significantly less data. Thus, HyperCBM offers a robust solution that maintains high task accuracy while providing model interpretability, which is especially valuable in label-constrained scenarios.

Table 2 ablates HyperCBM against leading semi-supervised baselines to verify its architectural advantages. The results yield two central conclusions: 1) HyperCBM consistently outperforms existing methods in low-data regimes, and 2) its performance advantage often scales with increased data availability. Even with only 10% of labels, HyperCBM establishes a substantial performance margin. On the PAS dataset, it leads the next-best method by +3.11% in class accuracy, while on

Table 3: Ablation study on PAS and BrEaST datasets (labeled ratio = 0.1). Best results are in **bold**, second-best are underlined.

HECRL	HIDP	*PAS*				*BrEaST*			
		Concept Acc.	Class Acc.	Concept AUC	Class AUC	Concept Acc.	Class Acc.	Concept AUC	Class AUC
-	-	73.64 ± 2.77	72.00 ± 2.75	55.82 ± 3.47	88.19 ± 1.25	69.19 ± 3.52	**65.49 ± 6.28**	50.12 ± 4.42	69.09 ± 6.54
✓	-	80.42 ± 1.88	72.44 ± 11.47	63.16 ± 4.37	87.33 ± 6.07	71.60 ± 3.12	63.53 ± 5.90	55.41 ± 4.56	70.11 ± 9.48
-	✓	80.64 ± 0.67	75.70 ± 4.07	63.86 ± 2.21	**90.16 ± 1.39**	70.65 ± 3.48	63.14 ± 3.38	54.11 ± 2.56	67.31 ± 6.04
✓	✓	**81.61 ± 0.88**	**76.89 ± 3.85**	**64.43 ± 1.27**	88.40 ± 2.51	**72.49 ± 3.90**	65.49 ± 7.08	**56.66 ± 3.96**	**70.31 ± 9.70**

BrEaST, it achieves top performance across all four metrics. Crucially, this performance gap widens at the 40% label ratio, highlighting our model's superior scalability. For instance, its concept accuracy leads on PAS grows to +4.78%, and it surpasses all competitors across all metrics on the SkinCon dataset. The consistent outperformance and scalability strongly suggest that HyperCBM's architecture is fundamentally more effective at leveraging heterogeneous supervision signals. Unlike baselines whose gains may plateau, HyperCBM is more adept at integrating semi-supervised information, leading to more robust concept representations and superior task accuracy. This firmly establishes HyperCBM as a new state-of-the-art for semi-supervised concept-based learning.

4.3 Interpretability and Test-time Intervention

Clinical Interpretability We generate concept activation maps with Grad-CAM (Selvaraju et al., 2017) for PAS and BrEaST test images, rescaling the heat-map intensity by the corresponding concept scores. Two board-certified radiologists independently examined the overlays and agreed that the highlighted regions coincide with the intended medical concepts. The upper-left panel in Fig. 3 shows the raw ultrasound image; the remaining panels depict selected concepts together with their activation maps (warmer colours indicate stronger evidence). Across both datasets, the highlighted areas match the regions that clinicians rely on for diagnosis, underscoring the practical plausibility of the model's concept-level explanations.

Test-time Intervention To assess faithfulness, we perform test-time interventions by applying a series of confidence thresholds $\tau \in \{0.1, 0.2, 0.4, 0.6, 0.8, 1.0\}$ to the concept prediction scores: any concept whose score exceeds τ is forced to zero, following Wang et al. (2024). A smaller τ therefore eliminates a larger fraction of high-confidence concepts, enabling us to examine how sensitive the final diagnosis is to the removal of concept evidence. As shown in Fig. 4, intervention causes a nearly monotonic degradation in diagnosis: with the harshest setting ($\tau = 0.1$), **Class AUC** drops from 86.67 % to 70.21 % and **Class ACC** from 72.89 % to 56.44 %. Performance steadily recovers as τ increases, reaching 84.14 % AUC and 71.26 % ACC at $\tau = 0.8$, and returning to baseline when no concepts are masked ($\tau = 1.0$). The consistent decline confirms that the classifier relies strongly on the predicted concepts: the more evidence we suppress, the larger the performance loss.

4.4 Ablation Study

As shown in Tab. 3, we evaluate the contributions of HECRL and HIDP individually and in combination. The primary role of HECRL is to significantly enhance the model's concept-level understanding. Its inclusion leads to a dramatic concept accuracy boost of +6.78% on PAS and +2.41% on BrEaST. This demonstrates HECRL's remarkable effectiveness in refining concept representations by modeling their inter-dependencies. Complementing this, HIDP is instrumental in improving downstream task performance. It delivers a substantial +3.70% gain in class accuracy on PAS, showcasing its strength in effectively mapping the learned concepts to the final prediction. This highlights its crucial role in bridging concept and task learning. (***See Appendix for more ablations***)

5 Conclusion

In this work, we design a semi-supervised concept bottleneck model *HyperCBM* for interpretable medical image diagnosis, addressing the limitations of inter-concept dependency modeling and pseudo-label reliability. By introducing a concept-level hypergraph for structured reasoning and an image-level hypergraph for adaptive pseudo-labeling, our framework improves both interpretability and diagnostic accuracy. Experiments on our collected PAS dataset, breast ultrasound and dermoscopic image public datasets demonstrate superior interpretable performance over existing CBMs methods. This presents a powerful baseline for concept label-efficient medical image analysis and facilitates clinical applications.

REFERENCES

Angelica I Aviles-Rivero, Nicolas Papadakis, Ruoteng Li, Philip Sellars, Qingnan Fan, Robby T Tan, and Carola-Bibiane Schönlieb. Graphx net-chest x-ray classification under extreme minimal supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 504–512. Springer, 2019.

Danilo Avola, Luigi Cinque, Alessio Fagioli, Gianluca Foresti, and Alessio Mecca. Ultrasound medical imaging techniques: a survey. *ACM Computing Surveys (CSUR)*, 54(3):1–38, 2021.

Pietro Barbiero, Francesco Giannini, Gabriele Ciravegna, Michelangelo Diligenti, and Giuseppe Marra. Relational concept bottleneck models. *Advances in Neural Information Processing Systems*, 37:77663–77685, 2024.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Giuseppe Cali, Francesco Forlani, Cristoph Lees, I Timor-Tritsch, Josè Palacios-Jaraquemada, Andrea Dall'Asta, Amar Bhide, Maria Elena Flacco, Lamberto Manzoli, Francesco Labate, et al. Prenatal ultrasound staging system for placenta accreta spectrum disorders. *Ultrasound in Obstetrics & Gynecology*, 53(6):752–760, 2019.

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5948–5955, 2023.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.

Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C Thai, Kathleen Moore, Robert S Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical image analysis*, 79:102444, 2022.

Yanwen Chong, Yun Ding, Qing Yan, and Shaoming Pan. Graph-based semi-supervised learning: A review. *Neurocomputing*, 408:216–230, 2020.

Townim F Chowdhury, Vu Minh Hieu Phan, Kewen Liao, Minh-Son To, Yutong Xie, Anton van den Hengel, Johan W Verjans, and Zhibin Liao. Adacbm: An adaptive concept bottleneck model for explainable and accurate diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 35–45. Springer, 2024.

Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Y Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. *Advances in Neural Information Processing Systems*, 35:18157–18167, 2022.

Xiaolong Deng, Huisi Wu, Runhao Zeng, and Jing Qin. Memsam: taming segment anything model for echocardiography video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9622–9631, 2024.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35:21400–21413, 2022.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3558–3565, 2019.

Yifan Feng, Jiangang Huang, Shaoyi Du, Shihui Ying, Jun-Hai Yong, Yipeng Li, Guiguang Ding, Rongrong Ji, and Yue Gao. Hyper-yolo: When visual object detection meets hypergraph computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Joshua Fixelle. Hypergraph vision transformers: Images are more than nodes, more than edges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9751–9761, 2025.

Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, and Qionghai Dai. 3-d object retrieval and recognition with hypergraph analysis. *IEEE transactions on image processing*, 21(9):4290–4303, 2012.

Yue Gao, Yifan Feng, Shuyi Ji, and Rongrong Ji. Hgnn+: General hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3181–3199, 2022.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1820–1828, 2021.

Yunqi Gu, Tao Zhou, Yizhe Zhang, Yi Zhou, Kelei He, Chen Gong, and Huazhu Fu. Dual-scale enhanced and cross-generative consistency learning for semi-supervised medical image segmentation. *Pattern Recognition*, 158:110962, 2025.

Yan Han, Peihao Wang, Souvik Kundu, Ying Ding, and Zhangyang Wang. Vision hgnn: An image is more than a graph of nodes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19878–19888, 2023.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35:23386–23397, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Lijie Hu, Tianhao Huang, Huanyi Xie, Chenyang Ren, Zhengyu Hu, Lu Yu, and Di Wang. Semi-supervised concept bottleneck models. *CoRR*, abs/2406.18992, 2024. URL https://doi.org/10.48550/arXiv.2406.18992.

Yuchi Huang, Qingshan Liu, and Dimitris Metaxas.] video object segmentation by hypergraph cut. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 1738–1745. IEEE, 2009.

Eric Jauniaux, Sally Collins, and Graham J Burton. Placenta accreta spectrum: pathophysiology and evidence-based anatomy for prenatal ultrasound imaging. *American journal of obstetrics and gynecology*, 218(1):75–87, 2018.

Reda Abdellah Kamraoui, Vinh-Thong Ta, Nicolas Papadakis, Fanny Compaire, José V Manjon, and Pierrick Coupé. Popcorn: Progressive pseudo-labeling with consistency regularization and neighboring. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pp. 373–382. Springer, 2021.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. *arXiv preprint arXiv:2306.01574*, 2023a.

Injae Kim, Jongha Kim, Joonmyung Choi, and Hyunwoo J Kim. Concept bottleneck with visual concept filtering for explainable medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 225–233. Springer, 2023b.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Songning Lai, Lijie Hu, Junxiao Wang, Laure Berti-Equille, and Di Wang. Faithful vision-language interpretation via concept bottleneck models. In *The Twelfth International Conference on Learning Representations*, 2023.

Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020a.

Yuexiang Li, Jiawei Chen, Xinpeng Xie, Kai Ma, and Yefeng Zheng. Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pp. 614–623. Springer, 2020b.

Shaolei Liu, Siqi Yin, Linhao Qu, and Manning Wang. Reducing domain gap in frequency and spatial domain for cross-modality domain adaptation on medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1719–1727, 2023.

Tianming Liu, Eliot Siegel, and Dinggang Shen. Deep learning and medical image analysis for covid-19 diagnosis and prediction. *Annual review of biomedical engineering*, 24(1):179–201, 2022.

Lucie Charlotte Magister, Dmitry Kazhdan, Vikash Singh, and Pietro Liò. Gcexplainer: Human-in-the-loop concept-based explanations for graph neural networks. *arXiv preprint arXiv:2107.11889*, 2021.

Elham Nasarian, Roohallah Alizadehsani, U Rajendra Acharya, and Kwok-Leung Tsui. Designing interpretable ml system to enhance trust in healthcare: A systematic review to proposed responsible clinician-ai-collaboration framework. *Information Fusion*, pp. 102412, 2024.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Winnie Pang, Xueyi Ke, Satoshi Tsutsui, and Bihan Wen. Integrating clinical knowledge into concept bottleneck models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 243–253. Springer, 2024.

Anna Pawłowska, Anna Ćwierz-Pieńkowska, Agnieszka Domalik, Dominika Jaguś, Piotr Kasprzak, Rafał Matkowski, Łukasz Fura, Andrzej Nowicki, and Norbert Żołek. Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 11(1):148, 2024.

Sandeep Reddy. Explainability and artificial intelligence in medicine. *The Lancet Digital Health*, 4 (4):e214–e215, 2022.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2 (3):e190043, 2020.

I Sarris, C Ioannou, P Chamberlain, E Ohuma, F Roseman, L Hoch, DG Altman, AT Papageorghiou, International Fetal, and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). Intra-and interobserver variability in fetal ultrasound measurements. *Ultrasound in obstetrics & gynecology*, 39(3):266–273, 2012.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Ivaxi Sheth and Samira Ebrahimi Kahou. Auxiliary losses for learning generalizable concept-based models. *Advances in Neural Information Processing Systems*, 36:26966–26990, 2023.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Zixing Song, Xiangli Yang, Zenglin Xu, and Irwin King. Graph-based semi-supervised learning: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 8174–8194, 2022.

Sakhinana Sagar Srinivas, Rajat Kumar Sarkar, Sreeja Gangasani, and Venkataramana Runkana. Vision hgnn: An electron-micrograph is worth hypergraph of hypernodes. *arXiv preprint arXiv:2408.11351*, 2024.

Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

Balagopal Unnikrishnan, Cuong Nguyen, Shafa Balaram, Chao Li, Chuan Sheng Foo, and Pavitra Krishnaswamy. Semi-supervised classification of radiology images with noteacher: A teacher that is not mean. *Medical Image Analysis*, 73:102148, 2021.

Hongmei Wang, Junlin Hou, and Hao Chen. Concept complement bottleneck model for interpretable medical image diagnosis. *arXiv preprint arXiv:2410.15446*, 2024.

Hanguang Xiao, Yangjian Wang, Shidong Xiong, Yanjun Ren, and Hongmin Zhang. Cuamt: A mri semi-supervised medical image segmentation framework based on contextual information and mixed uncertainty. *Computer Methods and Programs in Biomedicine*, pp. 108755, 2025.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

Xiaojin Zhu. *Semi-supervised learning with graphs*. Carnegie Mellon University, 2005.

A Appendix

This appendix provides supplementary materials to complement the main paper, including LLM usage declaration, detailed dataset descriptions and an in-depth hyperparameter analysis.

For the purpose of reproducibility and thorough peer review, the core module code in our project is anonymously available at the following link: **`https://anonymous.4open.science/r/HyperCBM-ICLR-Submission`**. The full code and dataset will be available once accepted.

Below, we outline the contents of the appendix:

- **Section 1** documents the usage of Large Language Models (LLMs) in the manuscript preparation, clarifying their role in language refinement while emphasizing that all scientific contributions originate from the authors.
- **Section 2** offers comprehensive descriptions of the three datasets employed in our study: the public benchmarks BrEaST and SkinCon, and our newly proposed Placenta Accreta Spectrum (PAS) dataset. This section elaborates on the data composition and details our hybrid concept annotation pipeline for the PAS dataset, which combines a Vision-Language Model with rigorous validation from clinical experts.
- **Section 2** presents a comprehensive sensitivity analysis for the critical hyperparameters, such as λ_1 and λ_2. It includes the experimental setup and a thorough discussion of the results, providing empirical justification for our chosen configuration and illustrating the inherent trade-off between classification performance and concept interpretability.

A.1 Section 1: LLM Usage

We employed Large Language Models (LLMs) to assist in the preparation of this manuscript, specifically for improving clarity, readability, and overall presentation. The LLM was used for tasks such as rephrasing sentences, checking grammar, and enhancing the fluency of the text. Importantly, the LLM was not involved in any aspect of the research design, methodology, data analysis, or conceptual development. All scientific ideas, experimental procedures, and analytical insights were solely conceived and carried out by the authors. The role of the LLM was strictly limited to language refinement. The authors retain complete responsibility for the manuscript's content and have ensured that all LLM-assisted edits comply with ethical standards and do not constitute plagiarism or scientific misconduct.

A.2 Section 2: Dataset Detailed Descriptions

This section details the three datasets utilized in our study. We first introduce two publicly available benchmark datasets, BrEaST (Pawłowska et al., 2024) and SkinCon (Daneshjou et al., 2022), which serve as a basis for comparison. We then provide a comprehensive description of one of our main contributions, the proposed Placenta Accreta Spectrum (PAS) dataset. A detailed list of information for each dataset can be found in Table 4. For experiments, we follow a consistent data-splitting strategy. The images for each dataset are randomly partitioned into training, validation, and test sets with a 7:1:2 ratio and preprocessed by taking a center crop to a uniform size of 224×224 pixels. To ensure robust evaluation, every experiment is repeated five times using fixed random seeds $(1, 2, 42, 2024, 2025)$, and we report the results as mean ± standard deviation.

A.2.1 Publicly Available Benchmark Datasets

BrEaST The BrEaST dataset (Pawłowska et al., 2024) is a collection of breast ultrasound images annotated with seven concepts derived from the BI-RADS descriptors. The full dataset comprises 256 images. Following the experimental setup in (Wang et al., 2024), we utilize the 254 images corresponding to Benign and Malignant diagnoses for our experiments.

SkinCon Our study utilizes the SkinCon dataset (Daneshjou et al., 2022), which originates from the Fitz17k collection of 3,691 images (Groh et al., 2021). From the 3,673 publicly available images, we curate our adopted dataset following the two-stage methodology of Wang et al. (2024). To ensure statistical robustness, we first distill the original 48 concepts into a list of 22, retaining only

Table 4: A detailed list of the concept labels, modality, and category counts for the BrEaST, SkinCon, and PAS datasets. Each concept represents a distinct visual or clinical feature for analysis.

Dataset	Used Concept List	Modality	Categories
BrEaST	Irregular SHape (IRS), Not Circumscribed Margin (NCM), Hyperechoic or Heterogeneous Echogenicity (HoHE), Posterior Features (PF), Hyperechoic Halo (HH), CALcifications (CAL), Skin Thickening (ST)	Ultrasound	Malignant Benign
SkinCon	PAPule (PAP), PLAque (PLA), PUStule (PUS), BULla (BUL), PATch (PAT), NODule (NOD), ULCer (ULC), CRUst (CRU), EROsion (ERO), ATRophy (ATR), EXUdate (EXU), TELangiectasia (TEL), SCALe (SCAL), SCAR (SCAR), FRIable (FRI), Dome-SHaped (DSH), Brown-Hyperpigmentation (BrH), White-Hypopigmentation (WhH), PURple (PUR), YELlow (YEL), BLAck (BLA), ERYthema (ERY)	Dermoscopic	Malignant Benign Non-neoplastic
PAS	**Placental Location:** Normal Location (NL), Low-lying (LL), Placenta Previa (PP) **Placental Thickness:** Thickness $<$ 30 mm ($T_{<30}$), Thickness 30–50 mm ($T_{30\text{-}50}$), Thickness $>$ 50 mm ($T_{>50}$) **Retroplacental Space:** Present (RPS_P), Absent (RPS_A), None (RPS_N) **Retroplacental Myometrium:** $>$ 1 mm ($MYO_{>1}$), $\leq$ 1 mm ($MYO_{\leq 1}$), Absence (MYO_A) **Bladder Line:** Normal (BL_N), Interrupted (BL_I), Absence with Bulge (BL_{AB}) **Cervical Morphology:** Normal (CX_N), Incomplete (CX_I), Disappeared (CX_D) **Retroplacental Flow:** None (RF_0), Normal (RF_N), Increased (RF_I), Numerous/Confluent (RF_{NC}) **Global Flow:** None (GF_0), Normal (GF_N), Increased (GF_I), Massive (GF_M) **C-Section History:** None (CS_0), One (CS_1), $\geq$2 ($CS_{\geq 2}$) **Cysts:** None ($CYST_0$), Small ($CYST_S$), Numerous ($CYST_N$) **Fluid:** None (FLD_0), Small (FLD_S), Numerous (FLD_N) **Vendor:** GE, Samsung, Canon, Mindray	Ultrasound	1: Normal 2: Placenta Accreta 3: Placenta Increta

those with a minimum frequency of 50 samples. Subsequently, we filter the image corpus, keeping only the images annotated with at least one of these 22 concepts. This process yields our final set of 3,205 images, with a class distribution of Non-neoplastic (71.67%), Malignant (14.95%), and Benign (13.39%).

The rationale for our curation process stems from the dataset's characteristic long-tail concept distribution. A detailed analysis reveals that a small set of high-frequency concepts, such as Erythema (present in 66.43% of images) and Plaque (60.97%), co-occurs with numerous low-frequency concepts (e.g., Bulla at 2.00%). Furthermore, the resulting dataset exhibits a rich multi-label nature, with each image annotated with 2.95 concepts on average, and the majority of samples (72.55%) containing 2 to 3 distinct concepts.

A.2.2 The Proposed PAS Dataset

To facilitate research in interpretable, concept-based diagnosis of Placenta Accreta Spectrum (PAS), we introduce a new, expertly curated dataset.

Data Source and Composition The PAS dataset was collected from one top-tier Asian hospital (Due to anonymity, we do not specify here). It consists of 671 ultrasound images from patients diagnosed with varying degrees of PAS severity. Each image in the dataset is accompanied by two types of labels:

- **Task Label** (y): A PAS severity level from one of three classes: Normal, Placenta Accreta, Placenta Increta.
- **Concept Labels** (c): A set of 45 fine-grained concept annotations that describe the underlying ultrasound characteristics, which are organized into six categories.

Concept Annotation Pipeline We employed a rigorous two-stage pipeline to ensure the quality and clinical relevance of the concept labels:

- **Automated Concept Extraction via Multi-Prompting**: We utilized HuatuoGPT-Vision (Chen et al., 2024), a publicly available Vision-Language Model, for initial concept extraction. To elicit comprehensive and multi-faceted information from each image, we developed a structured hierarchy of prompts that were systematically applied. These prompts were organized into four distinct categories:
  - *General Description* (e.g., ”Write a detailed description of the given image.”): To capture broad visual features and context.
  - *Medical Feature Summary* (e.g., ”What specific patterns in the image suggest the diagnosis of placenta accreta spectrum?”): To guide the model towards clinically significant findings.
  - *Clinical Scene Q&A* (e.g., ”What features in this ultrasound indicate placenta accreta spectrum, and how would you classify its severity?”): To probe the model's diagnostic reasoning capabilities.
  - *Targeted Concept Elicitation* (e.g., ”Identify the key medical terms and concepts related to this image.”): To directly extract relevant medical terminology.

  This multi-prompt strategy generated a rich corpus of textual descriptions for each image, forming the foundation for the subsequent validation stage.
- **Expert Validation and Standardization**: The text descriptions generated by the VLM were subsequently reviewed and validated by two board-certified obstetricians, each with over 5 years of experience in prenatal ultrasound. These experts filtered the extracted information and mapped it to the 45 predefined, standardized concept categories. This hybrid approach combines the scalability of large language models with the precision of domain expertise, yielding a high-quality dataset for building trustworthy AI models.

A.3 Section 3: Hyperparameter Sensitivity Analysis

A.3.1 Motivation

The hyperparameters λ_1 and λ_2 in our overall objective function (Equation 11) are critical for balancing the model's focus between three key objectives: task performance ($\mathcal{L}_{task}$), concept supervision on labeled data ($\mathcal{L}_c$), and feature-concept alignment on unlabeled data ($\mathcal{L}_{align}$). The choice of these weights directly influences the trade-off between final classification accuracy and the interpretability afforded by the learned concepts. An imbalance could lead to a model that performs well on the task but learns meaningless concepts, or vice versa.

A.3.2 Empirical Findings

Sensitivity Analysis Overview To validate our hyperparameter configuration and provide a deeper understanding of the model's behavior, we conducted a comprehensive sensitivity analysis on the PAS dataset, with results summarized in Table 5. This analysis investigates the impact of the concept loss weight, λ_1, and the semi-supervised alignment loss weight, λ_2, across labeled data ratios ranging from 1% to 80%. Our findings provide empirical justification for the operational point $(\lambda_1, \lambda_2) = (1.0, 0.1)$ and elucidate the trade-offs inherent in our semi-supervised, concept-based framework.

Effect of Concept Loss Weight (λ_1) We first examined λ_1, which controls the influence of the supervised concept loss, $\mathcal{L}_c$. Increasing λ_1 from 0.1 to 1.0 (while fixing $\lambda_2 = 0.1$) led to consistent improvements in concept-centric metrics (Concept ACC and AUC). This confirms that stronger supervision effectively guides the model toward accurate and interpretable representations aligned with ground-truth concepts. However, this gain came at the cost of a slight but consistent reduction in classification performance (Class ACC and AUC). More importantly, an excessively large weight ($\lambda_1 = 2.0$) sharply degraded both task performance and concept generalization, suggesting that over-constraining the model to the concept vocabulary impedes learning of complementary features. Thus, $\lambda_1 = 1.0$ emerges as a balanced choice that maximizes concept alignment without unduly sacrificing predictive accuracy.

Effect of Alignment Loss Weight (λ_2) Simultaneously, we analyzed the sensitivity to λ_2, the weight for our proposed semi-supervised alignment loss, $\mathcal{L}_{align}$. This component is paramount for leveraging unlabeled data. The analysis unequivocally demonstrates that model performance is highly responsive to this parameter. Across all labeled data ratios, all four performance metrics consistently peaked at $\lambda_2 = 0.1$. For instance, at a 5% labeled ratio, increasing λ_2 from 0.05 to 0.1 catapulted the Class ACC from 57.48% to 72.15%, underscoring the profound efficacy of the alignment loss. Conversely, further increasing λ_2 beyond this optimal point resulted in a steady performance decline. This behavior strongly suggests that while a modest alignment signal is highly beneficial for regularizing the model and improving generalization, an excessive weight probably introduces noise from the less reliable pseudo-labels, $\hat{\mathbf{c}}_{\text{pseudo}}$, thereby corrupting the learned feature space. This finding empirically validates our choice of $\lambda_2 = 0.1$ as the optimal weight to harness the potential of unlabeled data.

Impact of Labeled Data Ratios We further evaluated performance under varying labeled data ratios at the optimal hyperparameter setting. As expected, all metrics improved monotonically with more labeled data, with the most substantial gains observed in low-data regimes (1% to 20%). Beyond this point, performance began to plateau, indicating diminishing returns. This highlights both the robustness of our framework and the essential role of the semi-supervised component in ensuring competitive performance when labeled data is scarce.

Summary This exhaustive analysis provides principled justification for our hyperparameter configuration of $(\lambda_1, \lambda_2) = (1.0, 0.1)$ on the PAS dataset. It clarifies the interplay between direct concept supervision, semi-supervised signal alignment, and task-specific objectives, demonstrating that the chosen parameters achieve a robust balance between classification accuracy and concept interpretability.

Table 5: Sensitivity analysis of loss balancing weights on the PAS dataset. We vary λ_1 (concept loss) and λ_2 (alignment loss) under different labeled data ratios. Performance is reported as mean ± std over five runs. The configuration used in the main experiments is highlighted in **bold**; it shows the trade-off between task performance and concept interpretability.

Labeled Ratio	**Hyperparameters (λ_1, λ_2)**	**Concept ACC**	**Class ACC**	**Concept AUC**	**Class AUC**
0.01	*Varying λ_1 with fixed $\lambda_2 = 0.1$*				
	(0.1, 0.1)	70.34 ± 1.20	81.48 ± 2.65	53.23 ± 1.15	92.18 ± 1.48
	(0.5, 0.1)	69.58 ± 2.73	61.04 ± 8.53	53.03 ± 1.34	78.34 ± 7.29
	(1, 0.1)	70.21 ± 1.44	57.33 ± 9.88	52.67 ± 1.97	76.21 ± 7.48
	(2, 0.1)	69.49 ± 1.74	48.30 ± 4.67	50.37 ± 1.02	54.04 ± 6.30
	Varying λ_2 with fixed $\lambda_1 = 1$				
	(1, 0.05)	68.35 ± 2.09	48.30 ± 6.15	51.02 ± 0.85	59.59 ± 4.96
	(1, 0.1)	70.21 ± 1.44	57.33 ± 9.88	52.67 ± 1.97	76.21 ± 7.48
	(1, 0.2)	68.96 ± 2.70	48.00 ± 2.67	50.63 ± 0.78	57.94 ± 9.43
	(1, 0.4)	70.29 ± 1.44	52.30 ± 4.31	50.67 ± 1.26	64.24 ± 6.72
	(1, 0.6)	69.98 ± 1.63	49.63 ± 3.89	50.72 ± 1.30	60.31 ± 10.28
0.05	*Varying λ_1 with fixed $\lambda_2 = 0.1$*				
	(0.1, 0.1)	74.36 ± 2.29	79.85 ± 1.09	57.67 ± 1.32	92.53 ± 0.95
	(0.5, 0.1)	77.86 ± 1.28	75.41 ± 3.61	61.06 ± 1.31	88.70 ± 1.97
	(1, 0.1)	77.54 ± 0.84	72.15 ± 2.08	60.18 ± 1.79	86.25 ± 1.99
	(2, 0.1)	72.60 ± 2.56	47.85 ± 3.65	52.08 ± 2.00	63.29 ± 5.63
	Varying λ_2 with fixed $\lambda_1 = 1$				
	(1, 0.05)	74.48 ± 1.68	57.48 ± 4.81	56.83 ± 1.22	74.26 ± 3.58
	(1, 0.1)	77.54 ± 0.84	72.15 ± 2.08	60.18 ± 1.79	86.25 ± 1.99
	(1, 0.2)	76.33 ± 2.06	58.22 ± 7.05	58.39 ± 1.53	76.76 ± 3.52
	(1, 0.4)	73.79 ± 4.15	54.67 ± 7.37	53.01 ± 2.67	67.79 ± 10.26
	(1, 0.6)	71.73 ± 3.58	51.56 ± 4.97	51.87 ± 3.45	60.68 ± 10.72
0.1	*Varying λ_1 with fixed $\lambda_2 = 0.1$*				
	(0.1, 0.1)	75.18 ± 2.11	80.44 ± 2.59	58.21 ± 2.24	92.96 ± 0.96
	(0.5, 0.1)	81.76 ± 0.95	77.34 ± 1.00	65.00 ± 1.67	90.68 ± 1.77
	(1, 0.1)	81.61 ± 0.88	76.89 ± 3.85	64.43 ± 1.27	88.40 ± 2.51
	(2, 0.1)	76.89 ± 2.04	55.26 ± 3.85	56.63 ± 2.85	72.17 ± 4.30
	Varying λ_2 with fixed $\lambda_1 = 1$				
	(1, 0.05)	80.10 ± 1.08	71.55 ± 5.19	62.85 ± 1.45	83.99 ± 3.39
	(1, 0.1)	81.61 ± 0.88	76.89 ± 3.85	64.43 ± 1.27	88.40 ± 2.51
	(1, 0.2)	80.92 ± 1.24	69.04 ± 4.33	62.14 ± 1.59	83.62 ± 4.58
	(1, 0.4)	75.42 ± 3.67	57.48 ± 8.12	55.64 ± 4.70	74.63 ± 7.31
	(1, 0.6)	70.85 ± 3.64	50.37 ± 3.86	52.00 ± 4.00	62.08 ± 7.53
0.2	*Varying λ_1 with fixed $\lambda_2 = 0.1$*				
	(0.1, 0.1)	73.54 ± 2.69	81.63 ± 2.11	55.67 ± 3.39	92.91 ± 1.08
	(0.5, 0.1)	83.23 ± 0.47	80.30 ± 2.41	67.52 ± 1.22	92.47 ± 2.00
	(1, 0.1)	83.07 ± 0.70	76.59 ± 2.84	67.08 ± 2.05	90.57 ± 1.60
	(2, 0.1)	82.11 ± 0.96	72.44 ± 2.27	65.34 ± 2.17	84.85 ± 3.08
	Varying λ_2 with fixed $\lambda_1 = 1$				
	(1, 0.05)	83.09 ± 0.53	76.74 ± 4.48	66.67 ± 1.61	90.11 ± 2.19
	(1, 0.1)	83.07 ± 0.70	76.59 ± 2.84	67.08 ± 2.05	90.57 ± 1.60
	(1, 0.2)	83.86 ± 0.62	77.48 ± 1.20	67.27 ± 1.83	90.75 ± 1.53
	(1, 0.4)	80.19 ± 5.19	72.00 ± 10.54	63.05 ± 5.40	86.05 ± 5.84
	(1, 0.6)	75.08 ± 6.67	60.89 ± 15.19	56.15 ± 7.86	71.39 ± 16.71
0.4	*Varying λ_1 with fixed $\lambda_2 = 0.1$*				
	(0.1, 0.1)	73.42 ± 1.94	81.48 ± 2.10	55.37 ± 1.04	92.98 ± 0.53
	(0.5, 0.1)	84.15 ± 1.24	80.89 ± 2.54	68.52 ± 2.54	92.47 ± 1.77
	(1, 0.1)	84.19 ± 1.11	78.82 ± 4.18	68.20 ± 3.00	90.48 ± 3.03
	(2, 0.1)	79.75 ± 5.96	65.19 ± 13.14	62.96 ± 8.84	81.42 ± 10.58
	Varying λ_2 with fixed $\lambda_1 = 1$				
	(1, 0.05)	84.32 ± 1.67	77.19 ± 7.13	69.20 ± 4.69	89.64 ± 4.06
	(1, 0.1)	84.19 ± 1.11	78.82 ± 4.18	68.20 ± 3.00	90.48 ± 3.03
	(1, 0.2)	84.50 ± 1.65	76.74 ± 5.82	69.13 ± 4.11	89.76 ± 4.20
	(1, 0.4)	77.22 ± 6.87	66.96 ± 12.48	58.93 ± 9.03	82.00 ± 9.05
	(1, 0.6)	74.12 ± 6.22	63.11 ± 12.10	54.65 ± 8.88	71.79 ± 17.28
0.6	*Varying λ_1 with fixed $\lambda_2 = 0.1$*				
	(0.1, 0.1)	73.95 ± 2.24	80.45 ± 2.83	55.79 ± 2.26	92.16 ± 2.65
	(0.5, 0.1)	85.11 ± 0.79	80.30 ± 1.91	69.79 ± 3.05	92.08 ± 2.41
	(1, 0.1)	85.57 ± 0.83	80.15 ± 1.65	71.25 ± 1.91	93.33 ± 2.46
	(2, 0.1)	85.15 ± 1.25	76.15 ± 4.17	70.53 ± 2.66	89.51 ± 2.47
	Varying λ_2 with fixed $\lambda_1 = 1$				
	(1, 0.05)	85.58 ± 0.89	80.89 ± 0.98	70.84 ± 2.55	92.04 ± 1.20
	(1, 0.1)	85.57 ± 0.83	80.15 ± 1.65	71.25 ± 1.91	93.33 ± 2.46
	(1, 0.2)	84.85 ± 1.50	76.74 ± 4.15	69.39 ± 4.07	91.99 ± 2.44
	(1, 0.4)	77.55 ± 6.83	61.04 ± 13.19	58.01 ± 10.01	74.18 ± 12.55
	(1, 0.6)	73.15 ± 6.40	50.22 ± 20.51	54.71 ± 9.15	62.33 ± 20.30
0.8	*Varying λ_1 with fixed $\lambda_2 = 0.1$*				
	(0.1, 0.1)	72.34 ± 2.10	80.30 ± 3.43	53.88 ± 2.58	92.45 ± 0.93
	(0.5, 0.1)	85.90 ± 0.66	83.26 ± 2.55	72.58 ± 2.45	93.60 ± 1.93
	(1, 0.1)	85.57 ± 0.86	81.93 ± 3.49	71.89 ± 1.36	91.95 ± 1.44
	(2, 0.1)	85.19 ± 1.76	78.08 ± 5.79	70.66 ± 5.77	89.08 ± 3.72
	Varying λ_2 with fixed $\lambda_1 = 1$				
	(1, 0.05)	86.35 ± 0.74	81.48 ± 1.93	73.35 ± 2.06	91.92 ± 1.50
	(1, 0.1)	85.57 ± 0.86	81.93 ± 3.49	71.89 ± 1.36	91.95 ± 1.44
	(1, 0.2)	81.06 ± 6.64	69.93 ± 20.20	66.10 ± 9.84	82.54 ± 18.05
	(1, 0.4)	78.37 ± 8.93	62.22 ± 22.37	62.99 ± 10.74	74.90 ± 21.26
	(1, 0.6)	73.75 ± 8.33	58.22 ± 20.17	57.40 ± 10.99	69.35 ± 20.02