3 CALIN: Intergroup Confidence Alignment From Null-Input Calibration

To overcome calibration errors and biases under the FS-ICL setting and to ensure calibration fairness among subgroups, we propose **CALIN**, an *inference-time calibration* method that contains a bi-level procedure – from *population-level* to *subgroup-level*. The goal is to provide fair and reliable confidence without requiring an additional training/validation set or access to the MLLM's parameters.

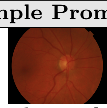
3.1 Notations

We assume that the predictive model is implemented by a pretrained frozen multimodal large language model $f_{\text{MLLM}}(\cdot)$ (e.g., GPT-4o and Gemini-1.5 [8,19]) that takes as input a set of *multimodal prompts* (image and text). We define a template φ that has fields for an image X, attributes A, and the label Y, though some may be left empty, to generate multimodal prompts. For example, $\varphi(X = \mathbf{x}, A = \text{Male}, Y = \text{Negative})$ is mapped to: "Does the fundus $\mathbf{x}$ of a male show glaucoma? Negative" (see Table. 1 for more examples). During inference, the model is provided with the multimodal prompt for the new query $\varphi(X = \mathbf{x}, A = a, \cdot)$ along with FS-ICL (few-shot) exemplars $\mathbf{D} := \{(X_i = \mathbf{x}_i, A_i = a_i, Y_i = y_i) | (X_i, A_i, Y_i) \in \mathcal{D}_{\text{fs}}\}$. The MLLM's predicted probability for $\hat{Y}$ being y given the inputs is denoted $\hat{p}_y(\mathbf{D}, \mathbf{x}, a)$ and estimated as follows:

$$\underbrace{\Pr\left[\hat{Y} = y \mid \mathbf{D}, X = \mathbf{x}, A = a\right]}_{\hat{p}_y(\mathbf{D}, \mathbf{x}, a)} = \frac{\Pr\left[\hat{T} = y \mid \mathbf{D}, X = \mathbf{x}, A = a\right]}{\sum_{y_j \in \mathcal{Y}} \Pr\left[\hat{T} = y_j \mid \mathbf{D}, X = \mathbf{x}, A = a\right]}. \quad (3)$$

Here, $\hat{T} = f_{\text{MLLM}}\big(\{\varphi(X_i, A_i, Y_i) | (X_i, A_i, Y_i) \in \mathcal{D}_{\text{fs}}\} \cup \{\varphi(X, A, \cdot)\}\big)$ is a random variable denoting the predicted next-token, and $\mathcal{Y} = \text{Val}(Y)$. We additionally define a vector $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a) \in \mathbb{R}^{|\mathcal{Y}|}$, where each dimension represents the probability of the prediction belonging to a specific class $y_j \in \mathcal{Y}$.

Table 1. Multimodal prompts φ under different inputs for fundus image classification. The left illustrates a datapoint containing the fundus image $X = \mathbf{x}$, the value of the attribute $A = \text{Male}$, and the label $Y = \text{Negative}$. The right illustrates the corresponding prompts. For cases $\varphi(\cdot, A, \cdot)$ and $\varphi(\cdot, \cdot, \cdot)$, we do not input the image to the MLLM.

Example Prompts φ		
	$\varphi(X, A, Y)$	*Does the fundus of a **male** show glaucoma? **Negative***
	$\varphi(X, A, \cdot)$	*Does the fundus of a **male** show glaucoma?*
	$\varphi(\cdot, A, \cdot)$	*Does an arbitrary fundus of a **male** show glaucoma?*
Male with **no glaucoma**	$\varphi(\cdot, \cdot, \cdot)$	*Does an arbitrary fundus show glaucoma?*

3.2 Bi-Level Confidence Calibration

The bi-level procedure used by CALIN can be intuitively thought of as first inferring the "amount of calibration" needed for the entire population (*population-level*), then inferring the "coarse" amount of calibration needed for each subgroup (*subgroup-level*). Information flows from the upper *population-level* to regularize the lower *subgroup-level* to provide an accurate and fair confidence calibration.

Population-Level Calibration $\mathscr{L}_1$**.** Inspired by the findings of language model's prediction bias presented in [27,6], CALIN first infers the "amount of calibration" for the entire population to avoid prediction bias under FS-ICL. In this work, the amount of population-level calibration is defined by a diagonal matrix $\mathbf{U} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ (we call it *calibration matrix* in this work), then the softmaxed linear transformation of $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a)$, determined by $\mathbf{U}$, is the $\mathscr{L}_1$ post-calibration confidence, given by $\bar{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a) = \operatorname{softmax}(\mathbf{U}\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a))$.

To determine $\mathbf{U}$ without the need of extra training/validation set, CALIN adopts a *multimodal null-input probing* technique. Specifically, we ensure that the predicted confidence $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a)$ is aligned with a uniform distribution when a null (or "content-free", "semantic-free" [27,15]) query $\varphi(\cdot, \cdot, \cdot)$ is fed into the MLLM. For a concrete binary classification example in Table. 1, when we neither provide the fundus image nor specify the sex of the patient, the MLLM's predicted confidence distribution should be uniform[5]. To this end, $\mathbf{U}$ is calculated based on the observed predicted confidence $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, \cdot)$ by the MLLM when we send null query $\varphi(\cdot, \cdot, \cdot)$ to it, given by $\mathbf{U} = (\operatorname{diag}(\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, \cdot)))^{-1}$.

Subgroup-Level Calibration $\mathscr{L}_2$**.** $\mathscr{L}_1$ improves confidence calibration over the entire population. To capture the potential variations across subgroups, we propose *subgroup-wise multimodal null-input probing* which aims to infer a set of calibration matrices $S := \{\mathbf{S}_a | a \in \mathcal{A}\}$ for $\mathscr{L}_2$ calibration. Each matrix in S focuses on calibrating one specific subgroup with sensitive attribute $A = a$. Borrowing from the intuition of multimodal null-input probing, subgroup-wise multimodal null-input probing finds S such that the predicted confidence given an attribute-conditioned null query $\varphi(\cdot, A = a, \cdot)$ is uniform for all subgroups. Specifically, we calculate them based on the observed predicted confidence $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, a)$ by the MLLM, given by $\mathbf{S}_a = (\operatorname{diag}(\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, a)))^{-1}$ for all $a \in \mathcal{A}$. Then, $\tilde{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a) = \operatorname{softmax}(\mathbf{S}_a\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a))$ is the $\mathscr{L}_2$ post-calibration confidence for any new query.

Regularizing $\mathscr{L}_2$ ***with*** $\mathscr{L}_1$**.** While $\mathscr{L}_2$ calibration aims to achieve subgroup level confidence alignment, relying solely on $\mathscr{L}_2$ may not guarantee accurate calibration. This is because the language model's inherent prompt bias [26,3] can lead to inaccurate and unstable estimation of calibration matrices, particularly since the $\mathscr{L}_2$'s probing prompt $\varphi(\cdot, A, \cdot)$ includes additional semantic information by conditioning on sensitive attributes. To mitigate this issue, we leverage $\mathscr{L}_1$ as

[5] We assume that it is impossible to identify the ground-truth label without observing the medical image $\mathbf{x}$.

a regularization mechanism, allowing the final calibration to capture subgroup variability and also penalizing anomalies. Specifically, we calculate a new set of calibration matrices $C := \{\mathbf{C}_a | a \in \mathcal{A}\}$ using exponential decay: When the estimated $\mathscr{L}_2$ calibration $\mathbf{S}_a$ extremely diverges (due to unstable estimation) from $\mathscr{L}_1$ calibration $\mathbf{U}$, the final calibration will be more aligned with $\mathscr{L}_1$, otherwise, the final calibration will be more aligned with $\mathscr{L}_2$. The decay rate is governed by $(\sqrt{\alpha+1})^{-1}$ where α is the maximum observed deviation across subgroups, calculated by $\alpha = \max_a\{\|\mathbf{S}_a\mathbf{i} - \mathbf{U}\mathbf{i}\|_\infty\}$ where $\|\cdot\|_\infty$ denotes the infinity-norm. The final calibration matrices are given by:

$$\mathbf{c}_a = \mathbf{U}\mathbf{i} + (\mathbf{S}_a\mathbf{i} - \mathbf{U}\mathbf{i}) \odot \exp\left(-(\sqrt{\alpha+1})^{-1} \cdot |\mathbf{S}_a\mathbf{i} - \mathbf{U}\mathbf{i}|\right), \tag{4}$$

$$\mathbf{C}_a = \mathrm{diag}(\mathbf{c}_a), \quad \forall a \in \mathcal{A}, \tag{5}$$

where $\mathbf{i} = \mathbf{1}_{|\mathcal{Y}|}$ is a vector with $|\mathcal{Y}|$ ones, $\odot$ denotes the element-wise product. We obtain the post-calibration confidence $\check{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a) = \mathrm{softmax}\left(\mathbf{C}_a \hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a)\right)$. Given the denoted vector construction $\check{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}, a) = [\check{p}_{y_j}(\mathbf{D}, \mathbf{x}, a) | y_j \in \mathcal{Y}]$ we can get the adjusted predicted label $\check{y} = \arg\max_{y_j \in \mathcal{Y}} \left\{\check{p}_{y_j}(\mathbf{D}, \mathbf{x}, a)\right\}$. The algorithm of CALIN is shown in Algorithm 1.

Algorithm 1 CALIN for Fair Confidence Calibration Under FS-ICL

Require: Few-shot $\mathbf{D}$, model f_{MLLM}, prompt template φ, demographic values $\mathcal{A}$
Ensure: Calibration matrices C
Compute $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, \cdot)$ using (3) with f_{MLLM}, $\mathbf{D}$, $\varphi(\cdot, \cdot, \cdot)$
Compute $\mathbf{U} = \left(\mathrm{diag}\left(\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, \cdot)\right)\right)^{-1}$ *#Population-Level#*
for a in $\mathcal{A}$ **do**
    Compute $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, a)$ using (3) with f_{MLLM}, $\mathbf{D}$, $\varphi(\cdot, A = a, \cdot)$
    Compute $\mathbf{S}_a = \left(\mathrm{diag}\left(\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \cdot, a)\right)\right)^{-1}$ *#Subgroup-Level#*
end for
Compute $\mathbf{C}_a$ using (4) and (5) with $\mathbf{U}$ and $\mathbf{S}_a$, add $\mathbf{C}_a$ to C. For all $a \in \mathcal{A}$
return C
Require: Few-shot $\mathbf{D}$, new query medical image $\mathbf{x}^*$, demographic value $a^* \in \mathcal{A}$, model f_{MLLM}, prompt template φ, calibration matrix $\mathbf{C}_{a^*} \in C$
Ensure: Adjusted prediction $\check{y}$ and its calibrated confidence $\check{p}$
Compute $\hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}^*, a^*)$ using (3) with f_{MLLM}, $\mathbf{D}$, $\varphi(X = \mathbf{x}^*, A = a^*, \cdot)$
Compute $\check{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}^*, a^*) = \mathrm{softmax}\left(\mathbf{C}_{a^*} \hat{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}^*, a^*)\right)$ *#Inference-Time#*
Assign vector elements $[\check{p}_{y_j}(\mathbf{D}, \mathbf{x}^*, a^*) | y_j \in \mathcal{Y}] = \check{\mathbf{p}}_{\mathcal{Y}}(\mathbf{D}, \mathbf{x}^*, a^*)$
return $\check{y} = \arg\max_{y_j \in \mathcal{Y}} \left\{\check{p}_{y_j}(\mathbf{D}, \mathbf{x}^*, a^*)\right\}$ and $\check{p} = \check{p}_{\check{y}}(\mathbf{D}, \mathbf{x}^*, a^*)$

4 Experiments and Results

Experiments are designed to showcase the effectiveness of CALIN in mitigating confidence calibration bias in MLLM under FS-ICL on 3 medical imaging datasets: (i) PAPILA [11], (ii) HAM10000 [21], (iii) MIMIC-CXR [10].