GLEN: Generalized Focal Loss Ensemble of Low-Rank Networks for Calibrated Visual Question Answering

Mahsa Mozaffari [1], **Hitesh Sapkota** [2*], **Qi Yu** [1]

[1]Rochester Institute of Technology,
[2]Amazon Inc.
mm3424@rit.edu, sapkoh@amazon.com, qi.yu@rit.edu

Abstract

Deep learning models with large-scale backbones have been increasingly adopted to tackle complex visual question answering (VQA) problems in real settings. While providing powerful learning capacities to handle the high-dimensional and multimodal VQA data, these models tend to suffer from the memorization effect leading to overconfident predictions. This can significantly limit their applicability in critical domains (*e.g.*, medicine, cyber-security, and public safety), where confidently wrong predictions may lead to severe consequences. In this work, we propose to perform novel low-rank network factorization, resulting in much better-calibrated networks. These low-rank factorized networks are then aggregated into an ensemble guided by a generalized focal loss to further improve the overall performance and calibration. The overall framework, referred to as the Generalized focal Loss Ensemble of low-rank Networks (GLEN), is an important step toward developing well-calibrated VQA models. We theoretically demonstrate that the generalized focal loss provides a more balanced bias-variance trade-off, which guarantees to lower the confidence of the incorrect predictions, resulting in improved calibration. Extensive experimentation conducted on benchmark datasets and comparison on various VQA models shows that GLEN leads to much better calibration over both in-distribution and out-of-distribution data without sacrificing the VQA accuracy.

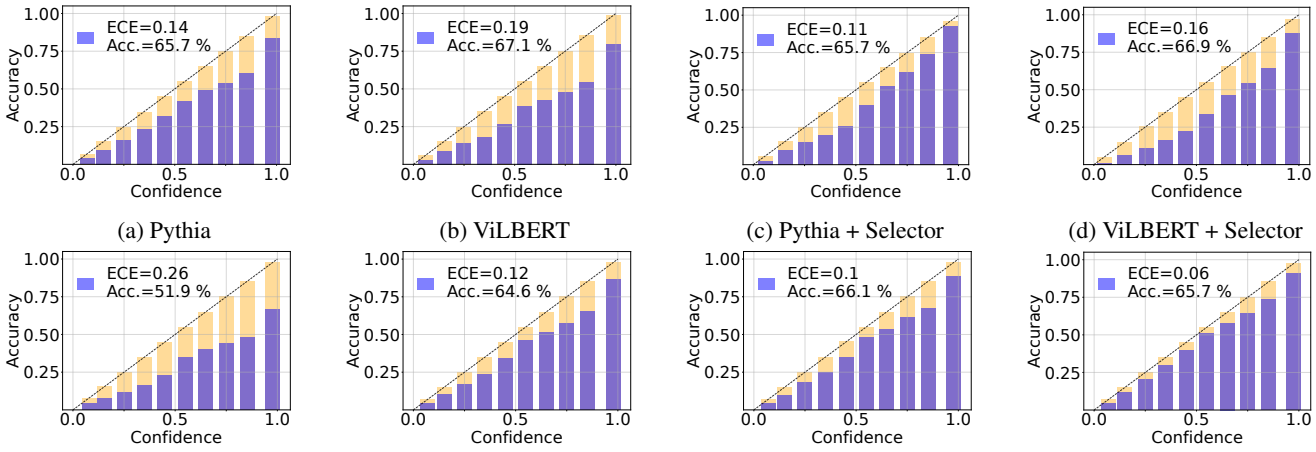
1 Introduction

Visual Question Answering (VQA) (Antol et al. 2015) has drawn significant attention due to its wide applicability in challenging real-world problems from diverse domains. Despite its wide applicability, VQA is inherently a challenging problem as it requires complex and common sense reasoning from multimodal data constituting images and natural language questions. To tackle the complex VQA problem, various methodologies have been developed (Schwenk et al. 2022; Lin et al. 2022; Gao et al. 2022; Qian et al. 2022; Vosoughi et al. 2023). However, most existing techniques are centered on enhancing VQA accuracy, without paying attention to the calibration of the model. Thus, these models are more likely to produce confidently wrong predictions due to the overfitting phenomenon stemming from the memorization effect of over-parametrized models. This issue undermines the reliability for VQA models, and negatively impacts user trust. For example as illustrated in Figure 1 both Pythia (Jiang et al. 2018) and ViLBERT (Lu et al. 2019) VQA models produce wrong predictions with a high confidence. As further demonstrated in Figures 2a, 2b, both models suffer from a poor calibration performance, despite having a decent VQA accuracy. Additionally, a better VQA accuracy does not necessarily ensure a better calibration. Such poor calibration behavior can severely limit the applicability of these VQA models in critical domains.

Figure 1: Illustration of over-confidence in VQA models: Pythia (Jiang et al. 2018) and ViLBERT (Lu et al. 2019) models, exhibiting incorrect answers with high confidence.

Some preliminary efforts have been devoted to achieving calibrated VQA model training. For example, (Whitehead et al. 2022; Dancette et al. 2023) introduces a trainable add-on component to a frozen VQA model, to estimate the confidence score associated with the VQA model's answer. They formulate VQA as a selective prediction problem, and the proposed techniques abstain from answering a question if the estimated confidence score by the selector falls below a threshold. However, there are two key limitations. First the training of the add-on selector depends on the VQA model, which may already be poorly calibrated, making the selector sub-optimal. Second, the selector needs to be trained on a standalone validation dataset, which significantly increases the annotation cost. As a result, those techniques still exhibit

*Work was completed during the PhD study at RIT, which is not related to the position at Amazon.



(a) Pythia (b) ViLBERT (c) Pythia + Selector (d) ViLBERT + Selector

(e) Factorized Early Layers (f) Factorized Final Layer (g) Ensemble of 3 LRF VQAs (h) Ours: GLEN

Figure 2: (a)-(b) The Expected Calibration Error (ECE) plots for Pythia and ViLBERT, indicating a high calibration error, despite a decent accuracy; (c)-(d) ECE plots of the Pythia and ViLBERT after applying a post-hoc calibration technique, Selector (Whitehead et al. 2022); (e)-(f) Comparison between varying effects of employing low-rank factorization to layers at different depths in Pythia (Jiang et al. 2018); Performance Comparison between (g) ensembling of low-rank final layer factorization of uniformly trained VQA models as in (f), and (h) our proposed GLEN.

poor calibration as shown in Figures 2c, 2d.

To address the poor calibration phenomenon of existing VQA models, one viable solution is to reduce the overall capacity of the overparameterized architecture by compressing the model weights through low-rank matrix/tensor factorization. However, compressing a complex deep architecture without care can hinder the model's representation ability leading to a poor prediction performance. This is because different layers in the network may play distinct roles in the learning process. Existing empirical observations and theoretical evidence suggest that earlier layers are responsible for learning general features and later ones focus on learning task-specific features, which may include noises resulting from spurious correlations or other sources (Kornblith et al. 2019; Yosinski et al. 2014; Allen-Zhu and Li 2022). These insights imply the potential to improve the calibration without compromising the predictive performance by keeping earlier layers dense while compressing the later layers that are primarily responsible for the overfitting phenomenon. This can be validated empirically as shown in Figures 2e, 2f wherein compression of the earlier layers leads to significant degradation of the model performance as it hinders the feature learning capability of the VQA model. In contrast, the compression of the final layer effectively enhances the calibration without causing much accuracy degradation.

To compensate the potential loss of modeling capability of the low-rank factorized (LRF) networks, one promising direction is to augment them using the deep-ensemble techniques with theoretically justified performance improvement (Allen-Zhu and Li 2023). As shown in Figure 2g, building an ensemble of three LRF VQA models shows an improved accuracy. Nevertheless, the overall ECE score is only improved by 2% and the ensemble still exhibits severe overfitting. For example, when the predicted confidence reaches over 80%, the accuracy is still lower than 70%. This issue arises because individual VQA models, trained in a similar manner from the same data distribution, lack diversity and exhibit strong correlations. As such, the ensemble potentially inherits the behavior (*e.g.* poor calibration) of individual VQA models. Therefore, it is important to ensure diversity among the individual models so that they can effectively complement each other during the ensemble phase.

Inspired by the above observations, in this paper, we propose a **G***eneralized Focal* **L***oss* **E***nsemble of Low-Rank* **N***etworks (*GLEN*)* framework which aims to produce well-calibrated VQA models without compromising the prediction performance. To achieve better calibration, we leverage the idea of *generalized focal loss (GFL)* that helps to diversify the LRF networks. GFL allows the learning of the LRF networks through different parts of data with distinct levels of difficulty and therefore enhances the diversity. Figure 3 shows that by coupling the training of LRF with the GFL, it leads to three networks (shown in dashed red) with diverse calibration behaviors. In contrast, without the GFL, the three LRF-Vanilla models (shown in dashed blue) are largely similar to each other. As a result, the GFL ensemble (shown in solid red) is able to achieve a better calibrated model than the standard ensemble (shown in solid blue): at the same accuracy, GLEN obtains a lower ECE score than the standard ensemble. This can be further confirmed by comparing Figures 2g and 2h. Our main contribution is fourfold:

- low-rank factorization on the later layers of the VQA network for improved calibration without hurting the feature representation capability.
- a generalized focal loss-based ensemble framework that combines multiple LRF networks that are diverse and complementary to each other to enhance the calibration.
- theoretical justification of the proposed technique by

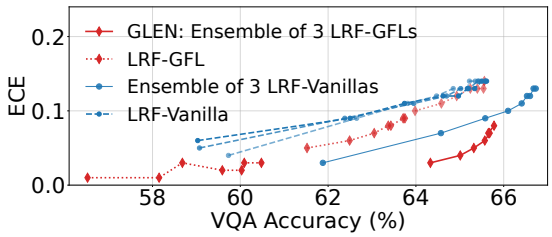


Figure 3: Effect of LRF with varying compression ratio. As the final layer of a VQA model gets more compact, ECE improves at the cost of slight accuracy degradation. Our technique consistently achieves a lower ECE at any accuracy level, as compared to adopting LRF on the original models with/without ensembling. This corroborates the effectiveness of the LRF combined with the GFL ensemble.

showing how the proposed generalized focal loss is equivalent to a bias-variance trade-off loss to guarantee the reduction of confidently wrong predictions.

- extensive experimentation to assess the effectiveness of the proposed technique in terms of in-domain and out-of-domain performance in VQA tasks.

2 Related Work

We give an overview of existing works most relevant to ours. Additional related works are covered in the Appendix.

Calibration in VQA models. The concept of reliable VQA is introduced by (Whitehead et al. 2022), which approaches it as a selective prediction problem. An add-on selection mechanism is leveraged to determine whether the model should provide an answer or abstain and the selection decision is based on the estimated confidence score. The selector component is trained separately from the main VQA model, which requires a good amount of labeled data. To address this, a training strategy is developed that trains both the VQA model and the selector on the same training dataset (Dancette et al. 2023). This is achieved by employing a distributed way for training the selector from N independent VQA models, each trained on $N-1$ splits. While these techniques help to improve the prediction reliability by abstaining from answering questions with low confidence, they do not fundamentally address the overfitting behavior of existing VQA models, which is the focus of the proposed work. Fig. 2 shows the selector still suffers from poor calibration.

Focal Loss and related models. Focal loss introduced by (Lin et al. 2017), is originally developed for object detection to address the class imbalance problem between the foreground and background. Later, (Mukhoti et al. 2020) explore its impact on the calibration of neural networks trained for classification. They demonstrate that focal loss enhances the calibration of neural networks, by having a regularization effect on the weights. AdaFocal (Ghosh, Schaaf, and Gormley 2022) emerges as a calibration-aware variant of focal loss, designed to dynamically adjust its hyperparameter for different sample groups. GFL (Li et al. 2020) extends the focal loss to the real-valued labels for object detection. In the domain of VQA, focal loss has been leveraged to tackle the dataset bias. For instance (Lao et al. 2021) utilized a focal loss to overcome language biases in VQA, implementing a strategy that reweights predictions made by a language-only branch. Our work proposes a generalized focal loss and then leverages it in a novel way to form a diverse ensemble of LRF networks for improved VQA calibration.

3 Methodology

Metrics for model calibration assessment. Let $\mathcal{D}_N = \{(\mathbf{v}_1, \mathbf{q}_1, \mathbf{a}_1), .., (\mathbf{v}_N, \mathbf{q}_N, \mathbf{a}_N)\}$ represent a dataset of N samples, where each input pair ($\mathbf{v}_n \in \mathcal{V}$, $\mathbf{q}_n \in \mathcal{Q}$) involves an image $\mathbf{v}_n$ and a question $\mathbf{q}_n$ and corresponding answer, denoted by $\mathbf{a}_n \in \mathcal{A}$, indicates the respective annotation. Further, we define $\mathcal{X} \equiv \mathcal{V} \times \mathcal{Q}$ to denote the input data space.

Unlike traditional classification problems, in the context of VQA, multiple ground-truth answers per image-question pair are available as multiple annotators annotate the same question. Thus, for each VQA task, the *accuracy (ACC)* could take a value in $[0, 1]$. Besides accuracy, *Expected Calibration Error (ECE)* (Naeini, Cooper, and Hauskrecht 2015) is commonly used to assess the calibration error between the estimated confidences and the actual accuracies. In many applications including the VQA task, over-confident wrong predictions directly impact the reliability of the VQA model and have a higher misleading effect, than under-confident correct predictions. Hence, it is crucial to reduce the overconfident prediction. Specifically, *Over-Confidence (OC)* metric (Mund, Triebel, and Cremers 2015), focuses on measuring the miscalibration within the wrong predictions: $OC = \mathbb{E}[\hat{p}|\hat{y} \neq y]$. There exists a trade-off between how often answers are abstained from, and the answer prediction error. The *Risk-Coverage* metric assesses this balance. Risk represents the average error on answered questions, while coverage measures the proportion of questions answered by the selective model. Given a desired risk threshold level R, the risk-coverage denoted by $C@R$ quantifies the maximum coverage achieved by the model while ensuring a minimum accuracy of $(1 - R)$ for answered questions, with higher C values being preferable. In critical applications, $C@R$ for lower risk threshold levels might be more critical to achieve. However, the overall selective prediction is summarized by *AUC*, which computes the total area under the $C@R$ curve. A comprehensive list of notations and further details on the above metrics are provided in the Appendix.

3.1 Overview of the Framework

Figure 4 shows the overall workflow of the `GLEN` framework, which consists of three key stages: 1) In the training stage, a diverse set of VQA models are trained with focal loss; 2) In the factorization stage, the final classification layers of the VQA models are factorized into a low-rank form; 3) Finally, in the ensembling stage, the low-rank VQA models are ensembled by aggregating their outputs.

3.2 Low-Rank Network Factorization

Overparameterization in neural networks is a well-recognized phenomenon that can benefit neural networks in

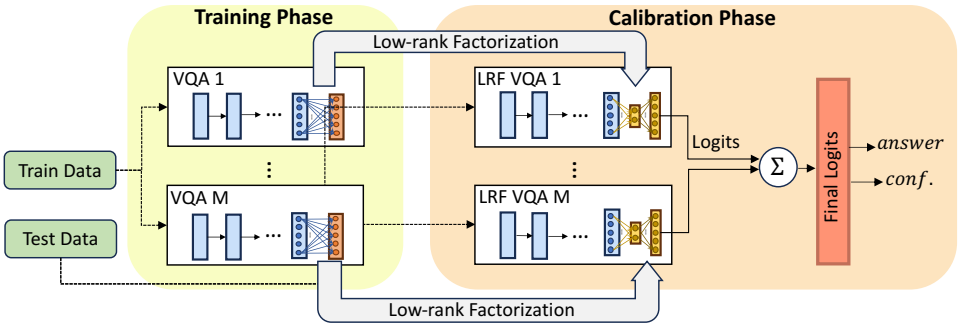


Figure 4: The overall workflow of generalized focal loss ensemble of low-rank networks

learning complex feature representation and capturing intricate features within data. The features captured at different depth levels vary in task-specificity, with the later layers learning task-specific and earlier layers capturing general features. Hence, to mitigate the poor calibration and overconfidence issue arising from an overparameterized architecture, we propose to conduct low-rank factorization on the final network layer, which not only reduces the model capacity at the final layer but also compresses the task-specific feature representations into a more compact form.

Specifically, we represent the final layer's weights by a low-rank weight matrix, effectively approximating the function performed by the layer. Assume that the weight matrix of the final layer is represented by $\mathbf{W} \in \mathbb{R}^{M \times C}$, which maps the penultimate representations $\mathbf{z} \in \mathbb{R}^{M}$ learned by the earlier layers, to probabilities across C classes, as $\mathbf{p} = \sigma(\mathbf{W}^{\top}\mathbf{z} + \mathbf{b})$, where $\mathbf{b} \in \mathbb{R}^{C}$, and $\sigma(.)$ respectively represent the layer bias, and the softmax function. Through low-rank factorization, the weight matrix $\mathbf{W}$ is approximated as $\mathbf{W} \approx \mathbf{U}\mathbf{V}$, where $\mathbf{U} \in \mathbb{R}^{M \times R}$, $\mathbf{V} \in \mathbb{R}^{R \times C}$, and R is the factorization rank. The layers within the neural networks typically exhibit a low-rank structure (Denton et al. 2014), which allows for approximating the weights with a factorization rank R, much smaller than M and C. The approximated function of the final layer is represented by

$$\mathbf{p} = \sigma(\mathbf{V}^{\top}\mathbf{s} + \mathbf{b}), \quad \mathbf{s} = \mathbf{U}^{\top}\mathbf{z}. \tag{1}$$

The functions represented in eq. (1) are effectively equivalent to two consecutive linear layers followed by a Softmax activation function, with the output space of the former layer being $\mathbf{s} \in \mathbb{R}^{R}$. Figure 5 visualizes the proposed low-rank network factorization. In this way, the original final layer's effective number of parameters reduces from MC to $R(M + C)$ parameters (excluding the bias term). Additionally, an implicit dimensionality reduction occurs at the task-specific features, accomplished by the new layers mapping the M-dimensional features $\mathbf{z}$ to an R-dimensional feature space by eq. (1), and subsequently mapping these intermediate R-dimensional features to output probabilities. The low approximation error between the original weight matrix and its low-rank form ensures that the model's predictive capabilities are mostly preserved.

We conduct empirical studies to verify the positive effects of applying low-rank factorization to the final layer

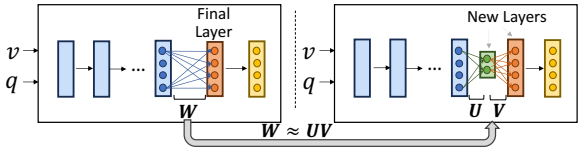


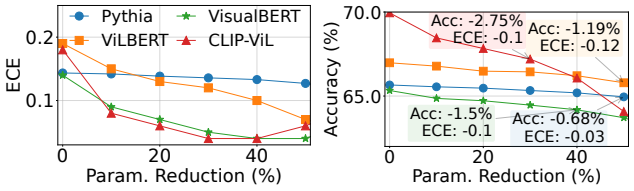
Figure 5: Low-rank factorization of the final layer

on the calibration of the VQA models. The technique is applied to four VQA models, namely, Pythia (Jiang et al. 2018), ViLBERT (Lu et al. 2019), VisualBERT (Li et al. 2019), and CLIP-ViL (Shen et al. 2021) with various factorization ranks. In particular, Figures 6a, 6b present the post-factorization performances in terms of the ECE and VQA accuracy, respectively, as a function of the ratio between layer parameter sizes, pre and post low-rank factorization stage. It's important to note that while approximating layer weights leads to some information loss, potentially impacting the accuracy of the model, our results highlight a promising trade-off: a slight drop in VQA accuracy, with an evident improvement in ECE, when compressing up to a certain ratio. Further empirical evidence, provided in Appendix F (Mozaffari, Sapkota, and Yu 2024), highlights the effectiveness of the final layer low-rank factorization on calibration and underscores the superiority of applying low-rank factorization to the final layer rather than to intermediate layers, in terms of the balance between VQA accuracy and ECE.

3.3 Generalized Focal Loss Ensemble

Performing low-rank network factorization helps to improve the calibration performance at the cost of slight degradation in the accuracy. In order to compensate this accuracy drop, we propose to construct an ensemble of the well-calibrated VQA models. Our empirical result (see Figure 3) shows that for individual VQA models trained in a similar manner from the same data distribution, they exhibit strong correlations. As such, the ensemble may inherit the limitation (*i.e.* poor calibration) of individual VQA models.

Therefore, it is crucial to ensure the diversity among the LRF networks in the ensemble model. We hypothesize that by enforcing the LRF networks to be trained through different data distributions, sufficient diversity can be achieved among different networks, making them complementary to each other during the ensemble process.



(a) (b)

Figure 6: ECE (6a) and VQA accuracy (6b) with respect to the final layer's parameter reduction ratio through low-rank factorization. The 0% reduction ratio corresponds to the performances of the original VQA models. As factorization rank decreases, the parameter size reduction ratio on the x-axis increases. In all models, lower factorization rank inflicts a small degradation in the VQA Accuracy while improving the ECE.

To achieve this, we propose a unique ensemble technique where each LRF network is trained using the generalized focal loss. Specifically, let $l(\mathbf{x}_n, \Theta)$ be the loss associated with the n^{th} data sample from the LRF network parameterized by Θ. Considering p_y^n being the output probability for the correct class y_n, the standard focal loss can be expressed as:

$$\mathcal{L}(\Theta)^{SFL} = \sum_{n=1}^{N} (1 - p_y^n)^{\gamma} l(\mathbf{x}_n, \Theta), \tag{2}$$

where γ controls the extent to which we want to give emphasis to the difficult samples. Depending on the γ value, the above expression instantiates different LRF networks. For instance, with $\gamma \to 0$, the above focal loss reduces to standard Expected Risk Minimization (ERM) loss. As such, the LRF network is trained by assigning equal weights to all samples. In this case, the model learns from the original data distribution. Further, with $\gamma \to \infty$, the above expression reduces to the maximum based loss where the LRF network focuses on the most difficult data sample and becomes unstable because it may attempt to learn from the noise. While the standard focal loss allows the LRF networks to focus on different data distributions depending on hyperparameter γ, the above expression lacks the statistical property to strictly justify the above phenomenon. Additionally, the exact interpretation of hyperparameter γ is not well understood. It also has a specific form and lacks the representation ability of different types of functions.

To overcome these shortcomings, we extend the focal loss to a general form:

$$\mathcal{L}(\Theta)^{GFL} = \sum_{n=1}^{N} w_n l(\mathbf{x}_n, \Theta), \tag{3}$$

where w_n indicates how much emphasis we want to impose on the n^{th} data points. It should be noted that depending on the weight distribution $\mathbf{w}$, the generalized focal loss puts emphasis on the difficult data samples. For example, by assigning equal weight on samples: $w_n = \frac{1}{N}; \forall n \in [N]$, the above focal loss reduces to putting equal emphasis on all data samples. In another extreme, by putting all weights to the sample with the highest loss, the above loss reduces to maximum based loss. Therefore, we would need a set that can define a distribution that can put an emphasis on difficult samples but to a different extent. Specifically, we define the following set for the weight distribution

$$\mathcal{W}_N := \left\{ \mathbf{w} \in \mathbb{R}^n, \mathbf{w}^\top \mathbb{1} = 1, 0 \leq \mathbf{w}, D_f \left(\mathbf{w} || \frac{\mathbb{1}}{N} \right) \leq \frac{\lambda}{N} \right\} \tag{4}$$

where D_f is the f-divergence metric measuring the distance between $\mathbf{w}$ and the uniform distribution. Based on the above constraint, we can find the optimal weight distribution $\mathbf{w}^*$ that maximizes the generalized focal loss. In this case, we perform the maximization in the generalized focal loss so that the model is forced to find the optimal weight $\mathbf{w}^*$ under the constraint set $\mathcal{W}_N$ that can favor more difficult samples by assigning higher weights. It should be noted that depending on the hyperparameter λ, our generalized focal loss can be reduced to the ERM loss (with $\lambda \to 0$) as well as maximum loss (with $\lambda \to \infty$). Therefore, by changing λ, we can systematically change the distribution over $\mathbf{w}$ to control the emphasis on difficult samples. Also, it is worth mentioning that, inspired by the work (Namkoong and Duchi 2017), the above generalized focal loss can also be expressed in terms of bias-variance trade-off loss which ensures the minimization of overconfident wrong predictions without reducing the prediction score for the correct cases. More formally, we present the following theorem:

Theorem 1 *Let X be a random variable representing a data sample, $\sigma^2 = Var[f(X)]$ and $Var_N[f(X)]$ denote the population and sample variance of $f(X)$, respectively, and D_f takes the form of χ^2-divergence. Let $|l(\mathbf{x}_n, \Theta) - \frac{1}{N}\sum_{i=1}^{N} l(\mathbf{x}_i, \Theta)| \leq K$ be the the upper bound of the absolute difference between data sample loss and average (ERM) loss. Then, under low-rank network factorization, with probability at least $1 - \exp\left(-\frac{Nt^2}{2K^2}\right)$, the generalized focal loss in eq. (3) can be represented by the bias-variance trade-off loss under the distribution set defined in eq. (4):*

$$\mathcal{L}(\Theta)^{GFL} = \frac{1}{N}\sum_{1}^{N} l(\mathbf{x}_n, \Theta) + C\sqrt{\frac{Var_N[l(X, \Theta)]}{N}}, \tag{5}$$

where $t = \left(1 - \frac{\sqrt{3}}{8}\right)\sigma$.

Remark. Because of the special design of our LRF networks, the equivalence between GFL and bias-variance trade-off loss is ensured with a high probability. This is because in the expression $1 - \exp\left(-\frac{Nt^2}{2K^2}\right)$, K indicates that the bound for the loss becomes small as the model restricts the output making the individual losses not deviate far from the average loss. As such, the probability of equivalence between the generalized focal loss and trade-off loss increases. In contrast, without LRF, the model may generate confident wrong predictions making some of the loss values very high. Consequently, the K-value remains large therefore reducing the probability for the equivalence.

The above bias-variance trade-off loss consists of two terms, where the first term indicates the bias in predicting each data sample, whereas the second term is related to the variance associated with the prediction. The optimization of the above loss tries to minimize both bias as well

as variance with the trade-off determined by hyperparameter C. It should be noted that the above optimization tries to make the model more generalizable by minimizing the variance and therefore our LRF networks become better in terms of avoiding false confident predictions. Furthermore, the first term ensures that we do not sacrifice the confidence of the correct predictions. Under our ensemble framework, we dedicate some LRF networks to focus on the representative (easy) samples by setting a smaller C value. Further, other LRF networks can focus on more difficult samples by setting a higher C value. As a result, these LRF networks become complementary to each other leading to better calibration for the final ensemble.

4 Experiments

We evaluate the performance of our proposed framework on multiple existing VQA models. For the evaluation, we follow the experimental setup of (Whitehead et al. 2022). To showcase the effectiveness of our technique, we report comparative quantitative results both in-distribution as well as out-of-distribution data. Further, we also conducted extensive qualitative analysis and ablation study to justify our approach. Due to limited space more results related to ablation study and qualitative analysis can be found in the Appendix.

Baselines. The "Baseline" in our tables indicates original VQA models, without any calibration mechanism. The maximum output probability is used as the confidence in the predicted answer. We also compare our approach to Temperature Scaling (TS) (Guo et al. 2017), Vector Scaling (Guo et al. 2017; Platt et al. 1999), and Selector (Whitehead et al. 2022) which are post-hoc calibration techniques. TS and Vector Scaling are standard calibration methods. The selector is trained as a regression task, where targets are the actual accuracy of the VQA model on the samples. All add-hoc calibration methods are trained on an additional validation data split, and hence require the availability of additional data, while the base VQA model layers are frozen.

Datasets. We experiment on the VQA-v2 (Goyal et al. 2017) and AdVQA (Sheng et al. 2021) datasets, respectively as in-distribution and out-of-distribution datasets. Additional experiments on VizWiz dataset (Gurari et al. 2018) are included in the Appendix due to space constraints. VQA-v2 dataset contains questions on the COCO image dataset, with 10 ground-truth answers per each question. The training split includes $443,757$ questions. As the ground-truth answers of the test split of VQA-v2 are not publicly available, we use the validation and test splits as provided by (Whitehead et al. 2022). The test split consists of 106k, and the validation split consists of 86k questions used. AdVQA dataset comprises of human-adversarial questions, on the same images as VQA-v2, crafted manually that are challenging to answer by models that are trained on VQA-v2. To assess the robustness against OOD examples, we evaluate models on a mixture of 90% ID, and 10% OOD questions from their corresponding test splits.

Backbone VQA models. We have extensively evaluated the performance of our presented framework by considering six different VQA architectures: LXMERT (Tan and Bansal 2019), Pythia (Jiang et al. 2018), CLIP-ViL (Shen et al. 2021), ViLBERT (Lu et al. 2019), and Visual-bert (Li et al. 2019), as well as the BEiT-3 (Wang et al. 2023) foundation model. Pythia (Jiang et al. 2018) is a bottom-up top-down model, and the winning model in the 2018 VQA challenge which leverages up-down attention mechanism (Anderson et al. 2018), and combines the representations of question and image by element-wise multiplication. CLIP-ViL (Shen et al. 2021) uses the Movie-MCAN architecture (Nguyen, Goswami, and Chen 2020) with the visual encoder of the CLIP (Radford et al. 2021) pre-training model. LXMERT (Tan and Bansal 2019), ViLBERT (Lu et al. 2019) and VisualBert (Li et al. 2019) are pre-training-based transformer architectures with an attention mechanism. BEiT-3 is a general-purpose vision-language model trained by masked-data modeling.

Evaluation Metrics. We evaluate using VQA accuracy, ECE, overconfidence (OC), and selective prediction performance measured through C@R at various risk levels: 1%, 5%, 10%, 20% as in (Whitehead et al. 2022). While our primary objective is to improve the calibration of VQA models, we also report VQA accuracy metrics into our evaluation in order to ensure that our model yields better calibration performance with comparable/better prediction performance.

4.1 Comparison Results

We evaluate the ID performance of the models and measure the generalizability of out-of-distribution scenarios. Table 1 offers a comparative analysis of our proposed method, GLEN against the baseline and other calibration techniques. It's evident that, across all VQA backbone models, GLEN achieves significantly lower ECE and OC scores than all competitors while achieving similar or better accuracies. Notably, our model exhibits an ECE improvement of approximately 11% over the second-best baseline on the VisualBERT model. Additionally, GLEN surpasses the baseline in terms of selective prediction, particularly in the $C@1$ metric accross all VQA architectures. Despite the competing calibration techniques benefit from additional training on a separate data split for confidence calibration, they fall short of GLEN or match GLEN in selective predictions, on LXMERT, Pythia, VisualBERT, and ViLBERT models. It's noteworthy that, a poorly calibrated model may still perform well in terms of $C@R$ metrics if the ordering of confidences matche their corresponding accuracies. The results underscores our technique's ability to significantly enhance model calibration, while also improving selective prediction performances. The effectiveness of GLEN in improving ECE is detailed for ViLBERT, VisualBERT and CLIP-ViL in Figure 8 in Appendix (Mozaffari, Sapkota, and Yu 2024).

The true value of enhanced calibration of models becomes apparent when dealing with OOD inputs that the models have not been trained to answer. In such cases, the models should naturally exhibit lower confidence in answering these inputs. As corroborated by Table 1, our method consistently achieves better calibration through lower ECE and OC, and outperforms or is comparable to the baselines in terms of selective prediction metrics. Particularly, considering the ViLBERT model, the proposed model has an improvement over

	Model	ID							ID+OOD						
		Acc.↑	ECE↓	OC↓	AUC↓	C@1↑	C@5↑	C@10↑	C@20↑	Acc.↑	ECE↓	OC↓	AUC↓	C@1↑	C@5↑	C@10↑	C@20↑
LXMERT	Baseline	**73.06**	0.14	0.56	8.74	12.89	42.54	62.63	88.73	**69.23**	0.16	0.56	11.24	2.55	31.55	51.75	81.28
	TS	**73.06**	0.13	0.54	8.74	12.89	42.54	62.63	88.73	**69.23**	0.15	0.54	11.24	2.55	31.55	51.75	81.28
	VectorScale	73.01	0.10	0.51	8.52	17.15	43.75	63.13	88.84	69.16	0.12	0.51	10.94	8.27	33.90	52.62	81.53
	Selector	**73.06**	0.09	0.52	8.34	19.42	45.24	**64.29**	88.68	**69.23**	0.11	0.52	**10.51**	**12.16**	**37.31**	**55.04**	**81.94**
	GLEN	72.93	**0.06**	**0.46**	**8.31**	**19.46**	**45.38**	64.03	**88.92**	69.07	**0.08**	**0.46**	10.74	10.88	36.01	53.14	81.61
Pythia	Baseline	65.67	0.14	0.51	13.56	6.12	26.28	42.85	72.67	62.01	0.16	0.51	16.24	3.34	20.22	34.96	63.14
	TS	65.67	0.10	0.46	13.56	6.12	26.28	42.85	72.67	62.01	0.12	0.46	16.24	3.34	20.22	34.96	63.14
	VectorScale	65.59	0.09	0.45	13.45	7.54	26.47	43.19	73.13	62.01	0.11	0.45	16.10	4.10	20.33	35.21	64.54
	Selector	65.67	0.11	0.50	13.34	8.34	27.48	43.51	73.48	62.01	0.13	0.49	15.87	**5.57**	21.90	36.48	64.98
	GLEN	**66.15**	**0.06**	**0.41**	**12.94**	**9.04**	**28.23**	**44.85**	**74.53**	**62.51**	**0.07**	**0.40**	**15.57**	4.76	**22.08**	**36.88**	**66.13**
CLIP-ViL	Baseline	69.95	0.18	0.58	10.77	6.78	34.01	54.23	82.57	**66.29**	0.20	0.58	13.39	1.50	24.87	44.48	73.33
	TS	69.95	0.16	0.55	10.77	6.78	34.01	54.23	82.57	**66.29**	0.18	0.56	13.39	1.50	24.87	44.48	73.33
	VectorScale	69.81	0.15	0.54	10.59	12.88	35.65	53.92	82.29	66.16	0.16	0.54	13.07	7.43	27.99	45.43	73.70
	Selector	69.95	0.13	0.55	**10.25**	**14.02**	**37.33**	**55.74**	82.82	**66.29**	0.15	0.54	**12.56**	**9.92**	**30.31**	**47.14**	**75.27**
	GLEN	**70.05**	**0.08**	**0.45**	10.46	10.32	35.65	54.61	**83.14**	**66.29**	**0.10**	**0.45**	13.06	7.01	27.00	43.84	75.10
ViLBERT	Baseline	**66.98**	0.19	0.57	13.00	1.70	24.54	45.10	75.93	63.20	0.21	0.57	15.92	0.00	16.22	35.17	65.83
	TS	**66.98**	0.17	0.54	13.00	1.70	24.54	45.10	75.93	63.20	0.18	0.54	15.92	0.00	16.22	35.17	65.83
	VectorScale	66.87	0.14	0.52	12.69	7.33	28.05	45.60	76.38	63.20	0.16	0.51	15.39	3.01	20.65	36.82	67.00
	Selector	**66.98**	0.15	0.55	12.25	**9.64**	**30.42**	47.65	77.01	63.20	0.16	0.53	14.81	**6.84**	**24.27**	38.71	68.72
	GLEN	66.90	**0.05**	**0.39**	**12.22**	9.23	29.60	**47.99**	**77.21**	**63.34**	**0.06**	**0.38**	**14.78**	3.83	22.56	**39.14**	**69.39**
VisualBERT	Baseline	64.92	0.14	0.50	14.06	6.06	24.75	41.22	71.31	61.39	0.16	0.50	16.69	3.35	18.48	33.91	61.56
	TS	64.92	0.14	0.49	14.06	6.06	24.75	41.22	71.31	61.39	15.02	48.66	16.69	3.35	18.48	33.91	61.56
	VectorScale	64.83	0.14	0.50	14.03	6.25	25.60	41.29	70.94	61.29	0.15	0.50	16.62	5.07	19.37	34.18	61.30
	Selector	64.92	0.15	0.54	13.79	6.78	26.53	42.43	71.90	61.39	0.16	0.53	16.19	**5.59**	21.40	35.92	63.71
	GLEN	**65.26**	**0.03**	**0.36**	**13.30**	**7.66**	**28.03**	**43.66**	**73.56**	**61.73**	**0.03**	**0.35**	**15.79**	5.32	**22.24**	**36.77**	**65.23**

Table 1: Performance comparison on the VQA-v2 test split (Whitehead et al. 2022) (ID) and AdVQA test split (OOD). In the ID+OOD setting, we test on a mixture of 90% ID and 10% OOD.

Model		Acc.↑	ECE↓	OC↓	AUC↓	C@1↑	C@5↑	C@10↑
BEiT-3	Baseline	74.68	0.09	0.55	7.78	14.86	47.62	66.89
	VectorScale	74.51	0.08	0.54	7.81	**18.01**	48.16	57.40
	GLEN	**74.95**	**0.02**	**0.43**	**7.43**	17.93	**48.66**	**68.36**

Table 2: Performance comparison for BEiT-v3 model.

10% on ECE, 13% on OC compared to the best competitor. This indicates that our model is extremely well calibrated and can be generalized well in the unknown environment where we encounter multiple out-of-distribution samples.

We also evaluate GLEN on a state-of-the-art large foundation model BEiT-3 model, fine-tuned on VQA-v2 in comparison with the standard VectorScale calibration. As shown in Table 2, the baseline BEiT-3 model shows superior accuracy and calibration performance as compared to the models presented in Table 1. Despite baseline BEiT-3 model's better calibration, our method further improves its calibration, reducing the ECE from 0.09 to 0.02, while simultaneously enhancing selective prediction performance. This underscores the robustness of our approach in effectively calibrating even highly accurate models, leading to more reliable and confident predictions across diverse scenarios.

4.2 Ablation Study

To further showcase the effectiveness of our proposed components, we have present the performances of different components on Pythia and VisualBERT models in table 3. As shown, without low-rank factorization and ensemble technique, the "Baseline" model exhibits poor calibration, as evidenced by its high ECE and OC scores. Constructing an ensemble on multiple dense VQA models somewhat improves the accuracy and ECE score. However, the ensemble tends to inherit the overfitting issues inherent in the individual models. Performing the low-rank matrix factorization in the later layers of models in the ensemble enhances the calibration performance. Yet, the LRF networks may still lack sufficient diversity, leading to sub-optimal performances. In contrast, by introducing diversity among the LRF networks by the generalized focal loss, GLEN significantly enhances both ECE and OC scores, without compromising accuracy. Further ablation studies are provided in the Appendix F (Mozaffari, Sapkota, and Yu 2024).

Model		Acc.↑	ECE↓	OC↓	AUC↓
Pythia	Baseline	65.67	0.14	0.51	13.56
	Ensemble	**66.75**	0.13	0.51	**12.62**
	LRF Ensemble	66.62	0.09	0.45	12.66
	GLEN	66.15	**0.06**	**0.41**	12.94
VisualBERT	Baseline	64.92	0.14	0.50	14.06
	Ensemble	**66.79**	0.12	0.49	**12.36**
	LRF Ensemble	66.41	0.09	0.43	14.37
	GLEN	65.26	**0.03**	**0.36**	13.30

Table 3: Ablation study demonstrating the effectiveness of GLEN components, on the VQA-v2 dataset.

5 Conclusion

By performing low-rank factorization on the VQA models, GLEN effectively alleviates the overconfidence issue while ensuring that the representation capabilities of VQA models are retained. Coupled with a generalized focal loss ensemble framework, the proposed technique ensures to train a diverse set of low-rank networks in the ensemble, each focusing on different data distributions, thereby complementing each other. This results in a better-calibrated ensemble model. Experimental results on both in-distribution and out-of-distribution scenarios using VQA-v2 and AdVQA datasets validate GLEN's effectiveness, showing its superiority in enhancing the VQA calibration.

References

Allen-Zhu, Z.; and Li, Y. 2022. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, 977–988. IEEE.

Allen-Zhu, Z.; and Li, Y. 2023. Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning. arXiv:2012.09816.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Dancette, C.; Whitehead, S.; Maheshwary, R.; Vedantam, R.; Scherer, S.; Chen, X.; Cord, M.; and Rohrbach, M. 2023. Improving Selective Visual Question Answering by Learning from Your Peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24049–24059.

Denton, E. L.; Zaremba, W.; Bruna, J.; LeCun, Y.; and Fergus, R. 2014. Exploiting linear structure within convolutional networks for efficient evaluation. *Advances in neural information processing systems*, 27.

Gao, F.; Ping, Q.; Thattai, G.; Reganti, A.; Wu, Y. N.; and Natarajan, P. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5067–5077.

Ghosh, A.; Schaaf, T.; and Gormley, M. 2022. AdaFocal: Calibration-aware Adaptive Focal Loss. *Advances in Neural Information Processing Systems*, 35: 1583–1595.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3608–3617.

Jiang, Y.; Natarajan, V.; Chen, X.; Rohrbach, M.; Batra, D.; and Parikh, D. 2018. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of neural network representations revisited. In *International conference on machine learning*, 3519–3529. PMLR.

Lao, M.; Guo, Y.; Liu, Y.; and Lew, M. S. 2021. A language prior based focal loss for visual question answering. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; and Yang, J. 2020. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33: 21002–21012.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, Y.; Xie, Y.; Chen, D.; Xu, Y.; Zhu, C.; and Yuan, L. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35: 10560–10571.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Mozaffari, M.; Sapkota, H.; and Yu, Q. 2024. Appendix: GLEN: Generalized Focal Loss Ensemble of Low-Rank Networks for Calibrated Visual Question Answering. Accessed: Dec 18, 2024. Available at: https://github.com/ritmininglab/GLEN.

Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P.; and Dokania, P. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33: 15288–15299.

Mund, D.; Triebel, R.; and Cremers, D. 2015. Active online confidence boosting for efficient object classification. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 1367–1373. IEEE.

Naeini, M. P.; Cooper, G.; and Hauskrecht, M. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Namkoong, H.; and Duchi, J. C. 2017. Variance-based Regularization with Convex Objectives. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Nguyen, D.-K.; Goswami, V.; and Chen, X. 2020. Movie: Revisiting modulated convolutions for visual counting and beyond. *arXiv preprint arXiv:2004.11883*.

Platt, J.; et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3): 61–74.

Qian, T.; Chen, J.; Chen, S.; Wu, B.; and Jiang, Y.-G. 2022. Scene graph refinement network for visual question answering. *IEEE Transactions on Multimedia*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 146–162. Springer.

Shen, S.; Li, L. H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; and Keutzer, K. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Sheng, S.; Singh, A.; Goswami, V.; Magana, J.; Thrush, T.; Galuba, W.; Parikh, D.; and Kiela, D. 2021. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34: 20346–20359.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Vosoughi, A.; Deng, S.; Zhang, S.; Tian, Y.; Xu, C.; and Luo, J. 2023. Unveiling Cross Modality Bias in Visual Question Answering: A Causal View with Possible Worlds VQA. *arXiv preprint arXiv:2305.19664*.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; and Wei, F. 2023. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Whitehead, S.; Petryk, S.; Shakib, V.; Gonzalez, J.; Darrell, T.; Rohrbach, A.; and Rohrbach, M. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, 148–166. Springer.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.