Appendix

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

In the Appendix, we first summarize the major notations used in our paper in appendix A. We then provide the proof of Theorem 1 in appendix B. We present additional related work in appendix C and give the detailed definitions of important calibration metrics in appendix E. We present additional experimental details and results in appendix F. Finally, we discuss broader impacts, and limitations and future works, respectively in appendix G and appendix H.

A Summary of Notations

Table 4 summarizes the major notations used in our paper.

Symbol Group	Notation	Description
Dataset	$\mathcal{A}$	Answer set
	$\mathcal{V}$	Image set
	$\mathcal{Q}$	Question set
	$\mathcal{V} \times \mathcal{Q}$	Input set
	$\mathbf{x}_n \equiv (\mathbf{v}_n, \mathbf{q}_n)$	Input image-question pair
	C	Total number of classes
Generalized Focal Loss	D_f	f-divergence
	$l(\mathbf{x}, \Theta)$	per-sample loss
	γ	Focal loss exponent
	p_y^n	Output probability for n-th data sample associated with class y
Low Rank Factorization	$\mathbf{W}$	Fully connected layer weight matrix
	$\mathbf{U}, \mathbf{V}$	Low-rank factorization factors
	Θ	Parameter associated with the given neural network
	R	Factorization rank (scalar)
	$\mathbf{z}$	A vector indicating the penultimate representation
	$\mathbf{p}$	Output probability distribution
	$\mathbf{b}$	Bias used to transform $\mathbf{z}$ to p
	$\sigma(\cdot)$	Softmax function
	M	penultimate representation dimensionality
Theoretical Results	$\mathcal{L}(\Theta)^{GFL}$	Generalized focal loss
	$\mathcal{W}_N$	Weight distribution associated with GFL
	$D_f(P\|\|Q)$	f-divergence measure between two distributions P and Q
	λ	Hyperparameter defining radius of ball in GFL to restrict $\mathcal{W}_N$
	N	Number of input data points
	σ^2	Population variance
	Var_N	Sample variance associated with N data samples
	K	Upper bound for the loss
	C	Parameter balancing bias and variance in GFL loss

Table 4: Symbols with Descriptions

B Proof for Theorem 1

In this section, we show the proof for theorem 1. We have the following generalized focal loss

$$\mathcal{L}(\Theta)^{GFL} = \sum_{n=1}^{N} w_n l(\mathbf{x}_n, \Theta) \tag{6}$$

where the weight distribution can be defined based on the following set

$$\mathcal{W}_N := \left\{ \mathbf{w} \in \mathbb{R}^n, \mathbf{w}^\top \mathbb{1} = 1, 0 \leq \mathbf{w}, D_f\left(\mathbf{w} || \frac{\mathbb{1}}{N}\right) \leq \frac{\lambda}{N} \right\} \tag{7}$$

By leveraging the focal loss, we would like to focus on learning from the difficult samples. Therefore, we can convert the above generalized focal loss into the following:

$$\mathcal{L}(\Theta)^{GFL} = \max_{\mathbf{w}\in\mathcal{W}_N} \sum_{n=1}^{N} w_n l(\mathbf{x}_n, \Theta) \tag{8}$$

Let us assume that $D_f(P,Q)$ is χ^2-divergence and since Q follows a uniform distribution in our weight distribution set, $D_f(P,Q)$ reduces to the Euclidean distance. Now let us consider some of the terminologies used in the paper. The ERM loss, *i.e.* the unweighted sum of losses over N data points can be regarded as the mean of loss and can be defined as $\bar{l} = \frac{1}{N}\sum_{n=1}^{N} l(\mathbf{x}_n, \Theta)$. Let's denote the loss vector is $\mathbf{l} = (l(\mathbf{x}_1, \Theta), l(\mathbf{x}_2, \Theta), ...l(\mathbf{x}_N, \Theta))^\top$. Then the empirical variance of $l(X)$ can be written as

$$\text{Var}_N[l(X, \Theta) = \frac{1}{N}||\mathbf{l}||_2^2 - \bar{l}^2 = \frac{1}{N}||\mathbf{l} - \bar{l}\mathbb{1}||_2^2 \tag{9}$$

where X denotes a random input variable. Furthermore, let's consider $\mathbf{v} = \mathbf{w} - \frac{\mathbb{1}}{N}$, so that eq. (8) (without maximization) can be represented as

$$\mathbf{w}^\top\mathbf{l} = (\mathbf{v} + \frac{\mathbb{1}}{N})\mathbf{l} = \bar{l} + \mathbf{v}^\top\mathbf{l} = \bar{l} + \mathbf{v}^\top(\mathbf{l} - \bar{l}\mathbb{1}) \tag{10}$$

In the above equation, we added additional term $-\mathbf{v}^\top\bar{l}\mathbb{1}$ as $\mathbf{v}^\top\mathbb{1} = 0$. Using the above equation, the generalized focal loss optimization becomes

$$\max_{\mathbf{v}\in\mathcal{R}^N} \bar{l} + \mathbf{v}^\top(\mathbf{l} - \bar{l}\mathbb{1}) \; s.t., ||\mathbf{v}||_2^2 \leq \frac{\lambda}{N^2}, \mathbf{v}^\top\mathbb{1} = 0, \mathbf{v} \geq -\frac{1}{N} \tag{11}$$

where the first constraint is achieved through the χ^2-divergence. By using the Cauchy-Schwarz inequality which states that $\mathbf{v}^\top\mathbf{u} \leq ||\mathbf{v}||_2||\mathbf{u}||_2$, we have the following:

$$\mathbf{v}^\top(\mathbf{l} - \bar{l}\mathbb{1}) \leq \frac{\sqrt{\lambda}}{N}||\mathbf{l} - \bar{l}\mathbb{1}||_2 = \sqrt{\frac{\lambda\text{Var}_N[l(X, \Theta)]}{N}} \tag{12}$$

In the above expression, for the necessary and sufficient condition to hold the above equality, we need the following:

$$\mathbf{v}_i == \frac{\sqrt{\lambda}(l(\mathbf{x}_n, \Theta) - \bar{l})}{N||\mathbf{l} - \bar{l}\mathbb{1}||_2} = \frac{\sqrt{\lambda}(l(\mathbf{x}_n, \Theta) - \bar{l})}{N\sqrt{N\text{Var}_N[l(X, \Theta)]}} \tag{13}$$

The second constraint in our optimization, *i.e.* $\mathbf{v} \geq -\frac{1}{N}$ holds if and only if

$$\min_{n\in[N]} \frac{\sqrt{\lambda}(l(\mathbf{x}_n, \Theta) - \bar{l})}{N\sqrt{N\text{Var}_N[l(X, \Theta)]}} \geq -1 \tag{14}$$

Therefore, if the above inequality holds, we have the following:

$$\max_{\mathbf{w}\in\mathcal{W}_N} \mathbf{w}^\top\mathbf{l} = \bar{l} + \sqrt{\frac{\lambda\text{Var}_N[l(X, \Theta)]}{N}} \tag{15}$$

By substituting $C = \sqrt{\lambda}$, we can exactly retrieve the bias-variance trade-off loss shown in eq. (5). This completes the proof of the theorem. One of the essential components considered in this proof is the use of eq. (14). Therefore, we need to show that with high probability eq. (14) holds true. Let us assume that our output loss is bounded, *i.e.* $|l(\mathbf{x}_n, \Theta) - \bar{l}| \leq K$. Then, to satisfy eq. (14), we need the following

$$\frac{2\lambda K^2}{N\text{Var}_N[l(X, \Theta)]} \leq 1 \quad \text{or} \quad \text{Var}_N[l(X, \Theta)] \geq \frac{2\lambda K^2}{N} \tag{16}$$

Defining the event, $\epsilon_N := \left\{\text{Var}_N[l(X, \Theta)] \geq \frac{3}{64}\sigma^2\right\}$ with sufficiently large number of data samples in the training set, *i.e.* $N \geq \frac{4K^2\lambda}{\sigma^2}\max\{2\sigma, 11\}$), we have the following:

$$N \geq \frac{44\lambda K^2}{\sigma^2} \geq \frac{2\lambda K^2}{\text{Var}_N[l(X, \Theta)]} \tag{17}$$

The above equation implies that, if the event $\epsilon_N := \left\{\text{Var}_N[l(X, \Theta)] \geq \frac{3}{64}\sigma^2\right\}$ happens with a high probability, eq. (14) also holds true. Now, we just need to show that with a high probability $\epsilon_N := \left\{\text{Var}_N[l(X, \Theta)] \geq \frac{3}{64}\sigma^2\right\}$ happens. For this, we leverage the Maurer and Pontil theorem which can be expressed in the form of the following lemma:

Lemma 1 (Maurer and Pontil Theorem 10) *Let U be a random variable taking values in [0, K]. Let $\sigma^2 = Var[U]$ be the population variance and $Var_N[U] = \frac{1}{N}\sum_{m=1}^{N} U_n^2 - (\frac{1}{N}\sum_{n=1}^{N} U_i)^2$ be the sample variance of U, respectively. Then for for $N \geq 2$,*

$$p(Var_N[U] \leq \sigma - t) \cup P(V_N[U] \geq \sigma + t) \leq \exp\left(\frac{-Nt^2}{2K^2}\right) \tag{18}$$

Setting $t = \left(1 - \frac{\sqrt{3}}{8}\right)\sigma$, the following holds

$$P(\epsilon_N) \geq 1 - \exp\left(-\frac{Nt^2}{2K^2}\right) \tag{19}$$

Remark. It should be noted that because the low-rank factorization over the later layers, our output variation would be bounded with respect to the input. As such, by having the non-ground truth predictions to be low, the loss will also be bounded making the K value small. As such, a smaller K will make the probability of $P(\epsilon_N)$ higher. Furthermore, in our context, the number of data samples N is also big, making $P(\epsilon_N)$ even higher. Therefore, with a very high probability, the generalized focal loss reduces to the bias-variance trade-off loss shown in eq. (5). This completes the proof.

C Additional Related Work

To complement the related work discussed in the main paper, we provide a high-level overview of general VQA models in this section. Recent research in Visual Question Answering (VQA) has seen substantial progress across multiple dimensions, reflecting a dynamic and evolving field. Attention mechanisms have been pivotal in advancing VQA, with various modalities and intra/inter-modal interactions explored (Anderson et al. 2018; Schwartz, Schwing, and Hazan 2017; Lu et al. 2018; Yu et al. 2019; Gao et al. 2019). Fusion techniques have been developed for an efficient and effective combination of visual and textual representations (Gao et al. 2016; Fukui et al. 2016; Kim et al. 2016; Ben-Younes et al. 2017; Schwartz, Schwing, and Hazan 2017). Knowledge-based VQA strategies have also contributed to the field by integrating the VQA models with external knowledge (Shao et al. 2023). Additionally, with recent advancements in the vision-language pertaining, VQA research has embraced the Transformer architecture and leveraged vision-language pretraining to approach or surpass human-level performance on benchmark datasets (Wang et al. 2021; Shen et al. 2021). Considering the unique challenges of VQA task, overcoming language biases and priors in VQA has gained traction, with specific techniques aimed at minimizing linguistic influences on model outputs (Cao and Li 2023; Lao et al. 2021). Despite having extensive literature on the VQA domain, the primary objective of existing techniques is to improve the prediction performance and therefore, very few of them are dedicated on improving the reliability/calibration of the model. In this paper, we focus on improving the calibration ability in the VQA tasks by leveraging generalized focal loss along with the strategic low rank factorization.

D Additional Details on Baseline Calibration Methods

Temperature Scaling (TS) (Guo et al. 2017) is a standard technique for calibrating neural networks. It involves learning a single temperature parameter, which is used to scale the logits of the model as $z' = \frac{z}{T}$, where z epresents the original logits, T is the learnable temperature, and z' are the scaled logits. The temperature parameter is optimized on a validation set using the Negative Log Likelihood (NLL) loss function.

Vector Scaling (Platt et al. 1999; Guo et al. 2017) extends Platt scaling to the multi-class setting as a standard post-hoc calibration method, which trains a logistic regression model on the logits (prior to softmax probabilities) of the neural network using a validation set. In all our experiments, the parameters of the main VQA model, while training a vector scaling.

Selector (Whitehead et al. 2022) is a post-hoc calibration method designed to enhance selective prediction performance in visual question answering (VQA). This method involves training a component that predicts a single probability output, which is optimized using cross-entropy loss between the predicted probability and the actual accuracy of the network's predictions. The selector component takes as input the answer logits, the question, the image, and the multimodal representations. It is trained using the validation split from the main VQA model's training process, along with an additional validation set.

For training the selector, on LXMERT, Pythia, ViLBERT, VisualBERT, and CLIP-ViL models, we followed the same approach as outlined in (Whitehead et al. 2022), i.e. selector with inputs from intermediate image, text, multimodal representations and answer logtis. For the BEiT-3 model, we trained a selector that inputs the answer logits and multimodal features from the final pooling layer before the classification head. For all models, during selector training, the parameters of the main VQA model were frozen, and only the selector was trained.

E Detailed Definitions of Metrics for Model Calibration Assessment

In this section, we provide the detailed definition for each of the metrics introduced in the main paper that can be leveraged for assessing the model's calibration performance.

Expected Calibration Error (ECE): ECE is calculated by dividing the N predictions into M equal bins according to their confidence scores. Within each bin B_m, the average accuracy and confidence are denoted by $\text{acc}(B_m)$ and $\text{conf}(B_m)$. Then, ECE is calculated as (Guo et al. 2017):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|,$$

where $|B_m|$ is the number of samples in the m-th bin. In the context of VQA, where there is more than a single ground-truth answer, ECE is measured with respect to the most frequent answer in the ground-truth annotations.

Over-Confidence (OC): OC measures the calibration error within the wrong predictions and a lower confidence score indicates better calibration. We empirically calculate OC by calculating the average confidence of the incorrect predictions, as:

$$OC = \frac{\sum_{n=1}^{N} \mathbb{I}(\hat{y} \neq y)\hat{p}}{\sum_{n=1}^{N} \mathbb{I}(\hat{y} \neq y)}.$$

Risk-Coverage: Risk-Coverage is introduced in the context of selective VQA. Assume $\mathbf{x} \in \mathcal{X}$ is an input pair of question and image. The selective VQA function is given by $h : \mathcal{X} \rightarrow \mathcal{A} \cup \{\emptyset\}$ and is defined as:

$$h_\tau(x) = \begin{cases} f(x) & \text{if } g_\tau(x) = 1 \\ \emptyset & \text{if } g_\tau(x) = 0 \end{cases}. \tag{20}$$

The selection mechanism is defined based on $g_\tau(x)$ which predicts a confidence score. A common metric for assessing selective prediction performance is the *Risk* and *Coverage*. Risk measures the error on the answered queries, while coverage is the proportion of the answered questions to the total number of questions. Risk and coverage are defined as follows:

$$C(g) = \frac{1}{N}\sum_{n=1}^{N} g_\tau(x), \quad R(f,g) = 1 - \frac{\sum_{n=1}^{N} Acc\big(f(x_n), y_n\big)\ g(x_n)}{\sum_{n=1}^{N} g(x_n)}, \tag{21}$$

where $Acc(f(x), y)$ is the VQA accuracy, defined by:

$$Acc\big(f(x), y\big) = \min\left(1, \frac{\#\text{ answers in } y \text{ matching } f(x)}{3}\right). \tag{22}$$

There exists a fundamental 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. While in critical applications. 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 compute the total area under the $C@R$ curve. A fundamental trade-off exists between risk and coverage. By varying the τ threshold, as the model attempts to increase coverage by answering more questions, the inherent risk associated with answering them naturally increases. This interplay between risk and coverage is quantified through the $C@R$ metric, denoting the maximum coverage attainable by a model, while ensuring that the associated risk remains below a specified threshold of R. The metric allows for a systematic assessment of a VQA's abstention performance.

F Additional Experimental Details and Results

In this section, we present the implementation details along with additional results to further justify the design choices and demonstrate the effectiveness of GELN.

F.1 Implementation Details

We use PyTorch for our experiments. For ViLBERT, VisualBERT, Pythia, and CLIP-ViL VQA models, we use their implementations as provided by the *MMF* (Singh et al. 2020) repository. For the LXMERT[1] and BEiT-3[2] models, we used their official source codes. To train the VQA backbones, the training hyperparameters of the networks given in repositories are used. Furthermore, for factorization of weight matrices, the PyTorch implementation of Singular Value Decomposition (SVD) is utilized. The SVD is then followed by the truncation of lower singular values to obtain low-rank weight matrix. For training of post-hoc calibration methods, including Selector, and Vector Scaling, the base VQA model is frozen.

[1]https://github.com/airsplay/lxmert

[2]https://github.com/microsoft/unilm/tree/master/beit3

For training LXMERT, Pythia, ViLBERT, VisualBERT, and CLIP-ViL, we utilized a single A6000 GPU, while the BEiT-3 model was trained on two A100 GPUs.

Throughout all experiments, we fix the number of models within ensembles to 3. The hyperparameters of our framework include the focal loss hyperparameter λ and low-rank factorization rank, which is set in a way to obtain a desired weight compression ratio. Throughout the paper, for simplicity we present ECE/Accuracy plots with respect to the weight size reduction ratio, rather than the factorization rank R. We do a grid search within a pre-defined range for λ and layer compression ratios. Within each low-rank ensemble configuration, we keep the factorization rank fixed for simplicity of grid search and identify hyperparameter combinations that yield the lowest ECE for which predictive performance is minimally or not affected compared to the original VQA.

Dataset	Architecture	Initial Network		LRF Network	
		Acc.	ECE	Acc.	ECE
CIFAR100	ResNet-101	75.7	0.12	74.9	0.03
	ResNet-152	75.3	0.08	74.7	0.02

Table 5: Effectiveness of LRF in calibrating ResNet models on CIFAR100

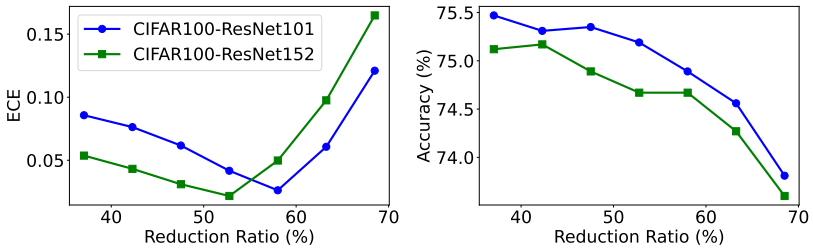


Figure 7: Accuracy and ECE performances vs. the reduction ratio in the final layer's parameter size by LRF

F.2 Reliability diagrams of GLEN VQA Models on VQA-v2 dataset

Figure 8 presents the reliability diagrams of calibrated models by GLEN, for ViLBERT, VisualBERT, CLIP-ViL, and LXMERT models.

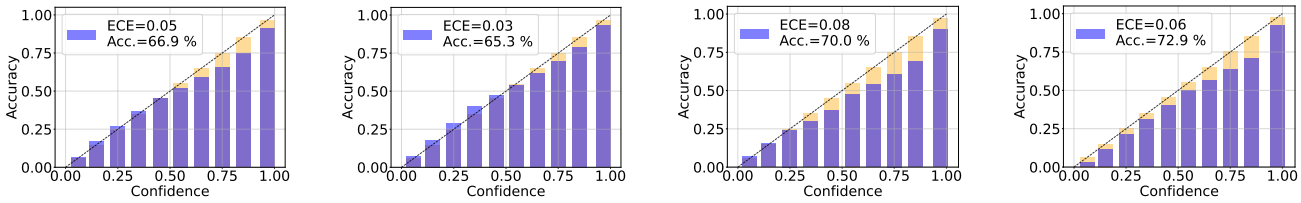
F.3 Effectiveness of LRF on Model Calibration: CIFAR-100 Image Classification

In this experiment, we demonstrate the calibrating effects of the LRF network on the standard image classification task. We test on several ResNet architectures including ResNet-101 and ResNet-152 using the CIFAR-100 dataset. We train models using the standard cross-entropy loss and apply low-rank factorization to the final linear layer of the trained models. In Table 5 we present a performance comparison of models before and after applying LRF in terms of classification accuracy and ECE. Notably, LRF on ResNet-101 reduces its ECE by 75% while largely maintaining the accuracy. This is with the number of parameters in the final layer reduced by $\sim 58\%$. fig. 7 demonstrates the accuracy and ECE with respect to the layer parameter reduction. With the increase of the reduction ratio, ECE first drops to a very low value and then starts to increase. This implies that it is necessary to keep certain capacity of the later layers in order for the model to learn well. On the other hand, with a 50-60% reduction ratio when the model become highly calibrated, the accuracy only drops less than 1%.

fig. 9a shows that the initial network exhibits the overconfidence issue and applying LRF helps to calibrate the network effectively. As factorization rank decreases (hence the reduction ratio increases), the network at first gets better calibrated as shown in figs. 9b and 9c. Then by further increasing the reduction ratio, the network will again become less calibrated, as shown in fig. 9d. In this case, the network becomes under-confident, leading to a higher ECE score. This behavior is consistent with the ECE performance shown in fig. 7.

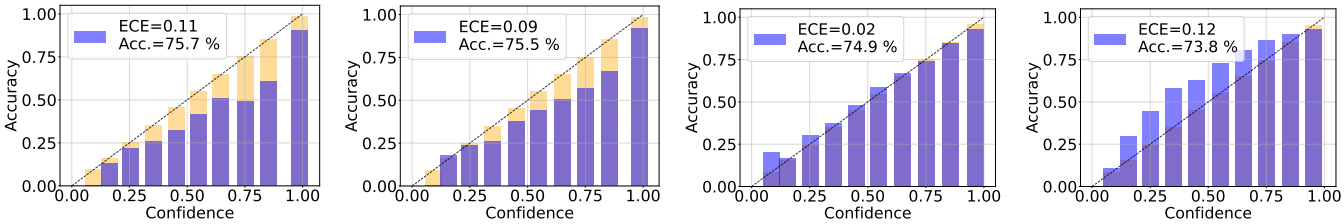
F.4 Effectiveness of Diverse LRF Networks

In this section, we provide complementary information on the performance evaluations presented in table 1. Particularly, we show how each LRF network within the ensemble performs and discuss their role in attaining a low ECE as well as high predictive and selective prediction performances. The results are reported based on the ViLBERT backbone and other backbones follow a similar trend. table 6 presents the performance of individual LRF networks that are ensembled together to form GLEN. The LRF networks differ in the GFL hyperparameters, *i.e.* λ in eq. (4), and the reduction ratio is set to 70%. As can be seen in fig. 10, LRF network 1 is under-confident as shown in fig. 10a with a low ECE and OC, as well as a low accuracy and



(a) (b) (c) (d)

Figure 8: GLEN in ViLBERT (8a), VisualBERT (8b), CLIP-ViL (8c), and LXMERT (8d)



(a) Initial Network (b) At 42% reduction (c) At 63% reduction (d) At 69% reduction

Figure 9: 9a ECE plot of the ResNet101 trained on CIFAR100; figs. 9b to 9d demonstrates ECE plots of low-rank factorized ResNet101, with reduction ratios 42%, 63%, 69%, respectively.

risk-coverage AUC. However, LRF networks 2 and 3, respectively shown in figs. 10b and 10c, are both over-confident, with latter being more over-confident than the other. LRF network 3 is trained with a GFL loss configuration that is close to the cross-entropy loss, hence performs similar to the baseline prior to LRF. The diversity among the LRF networks are evident by comparing their individual performances and ECE figures. The ensemble of 3 LRF networks as reported for GLEN, which not only has a low ECE (compared to the baselines), but achieves a high prediction accuracy (comparable to baseline) and enhances the selective prediction performances over the baselines.

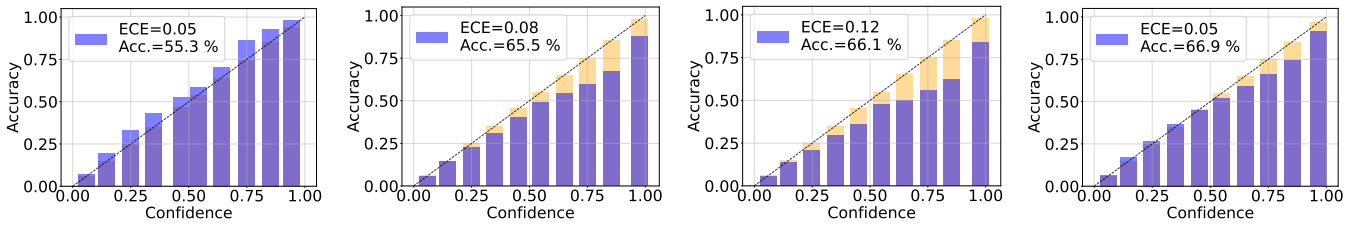
sub-network 1	sub-network 2	sub-network 3	Acc.	ECE	OC	AUC
✓	×	×	52.26	**0.05**	**0.30**	24.63
×	✓	×	65.46	0.08	0.43	13.42
×	×	✓	**66.11**	0.12	0.46	**13.18**
✓	×	✓	66.20	0.04	0.38	12.90
✓	✓	×	65.39	0.02	0.35	13.33
×	✓	✓	66.99	0.09	0.447	12.27
✓	✓	✓	66.90	0.05	0.39	12.22

Table 6: Performance of each LRF network within GLEN

F.5 Ablation Study

In this ablation study, first we compare the performance variance using low-rank factorization on the intermediate layer and final layer. This helps to highlight the importance of performing LRF on the later layer. We then show the contribution of each proposed component in the GLEN framework. We demonstrate the effect of the ensemble size within GLEN on the performances of the models. Finally, we study the impact of adopting diverse factorization ranks per each network within the ensemble in GLEN.

Low-Rank factorization on Penultimate vs. final layer of a network. In this study, we evaluate the impact of applying LRF to earlier layers compared to the final layer. Specifically, we factorize the second to the last layer of a VQA model. Our experiment shows that applying LRF to the final layer has a higher impact on enhancing the calibration. fig. 11 compares the calibration and VQA accuracy between the two cases: LRF on the final layer vs. penultimate layer in the Pythia, ViLBERT, VisualBERT, and CLIP-ViL models. In Pythia, applying LRF to penultimate slightly degrades ECE. In ViLBERT, VisualBERT and CLIP-ViL, while applying LRF to both improves the ECE, the final layer has a higher impact. This is evident by comparing the ECEs at the same VQA accuracy levels in both cases. At the same level of accuracy, the ECE of applying LRF to the last layer is lower than that of the penultimate layer. This observation aligns with our original intuition. By applying LRF to the final layer, we allow the model to learn important features, while the factorization of the final layer reduces the effect of noisy features through effective dimensionality reduction.



(a) Sub-network 1 (b) Sub-network 2 (c) Sub-network 3 (d) GLEN

Figure 10: Individual sub-networks of the GLEN for the ViLBERT backbone, the performance of which was reported in table 1.

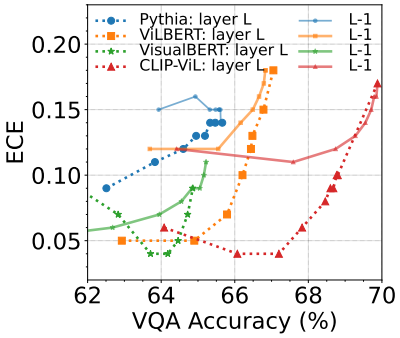


Figure 11: Effects of LRF on the final layer vs. the penultimate layer

Contribution of each component Table 7 shows the contribution of each proposed component. As shown, compared to the Baseline (*i.e.* the first row for each VQA model), by using LRF, the calibration improves. Furthermore, applying GFL based ensemble (without LRF) also improves the calibration. Applying ensemble on the top of LRF, we further improve the ECE. Finally, GLEN, which seamlessly integrates LRF and GFL based ensemble most effectively boosts the calibration and OC without sacrificing the accuracy.

Effect of Number of Ensemble Components This study assesses the impact of varying number of networks in the ensemble within the framework of GLEN. We analyzed the ensemble size's effects on calibration, overconfidence, accuracy, and selective prediction performances for ViLBERT and VisualBERT architectures. To ensure consistency with the findings in the rest of the paper, a same factorization rank that is consistent with the experiment in section 4.1 is used for all networks within ensembles, while varying the ensemble size from 2 to 5 networks. table 8 summarizes the performances. In the ViLBERT, increasing the ensemble size, leads to a reduction in ECE to 0.04, along with the improvements in accuracy and AUC metrics. In the case of VisualBERT, an ensemble of 2 networks achieves an ECE of 0.03, a result similar to that of a 3-network ensemble. However, expanding the ensemble to 4 networks further reduces the ECE to 0.02, thereby enhancing overconfidence, accuracy and AUC. It is noteworthy that increasing the number of components to 5, does not yield further gains in these metrics.

Uniform vs. Diverse Factorization Ranks In our main result, we adopt the same factorization rank across all LRF networks within an ensemble. In this set of ablation study, we explore the impact of applying different factorization ranks to individual LRF networks in the ensemble. To ensure fairness of comparison, for all VQA backbones, the same GFL setting as those presented in table 1 are used. Our findings as presented in table 9, demonstrates that non-uniform low-rank factorization can benefit GLEN in terms of either calibration or predictive performances. In the CLIP-ViL model, there is a notable enhancement in the accuracy and AUC, increasing accuracy to 70.23% from 70.05%, and reducing AUC to 10.27 from 10.46, while also improving ECE, reducing it to 0.07 from 0.08 obtained using the uniform factorization rank. Additionally, in the case of the ViLBERT model, the adoption of non-uniform ranks not only maintains the ECE, but also enhances the model's accuracy and selective prediction capabilities.

Effect of Post-Training Factorization on Calibration To demonstrate the calibration benefits of low-rank factorization (LRF) on the models, we conducted additional experiments by modifying the final layers of the VQA model to match the architecture of the LRF models reported in the paper. Specifically, we used the Pythia architecture for this analysis. The original Pythia model achieves an accuracy of 65.66 and an ECE of 0.14. As reported in Table 7 the LRF Pythia model achieves an accuracy of 65.37 and an ECE of 0.11, representing a 3% improvement in calibration compared to the original model. We trained an LRF model with the same bottleneck layer architecture from scratch. This model achieves an accuracy of 64.38 and an ECE of 0.14, which is the same as the original model. This indicates that designing a model with the same bottleneck layers

VQA Model	Model Choices			Acc.	ECE	OC
	GFL	LRF	Ensemble			
Pythia	×	×	×	65.66	0.14	0.51
	×	✓	×	65.37	0.11	0.46
	×	✓	✓	66.62	0.09	0.45
	✓	×	✓	66.30	0.10	0.47
	✓	✓	✓	66.15	0.06	0.41
CLIP-ViL	×	×	×	69.95	0.18	0.58
	×	✓	×	68.46	0.08	0.45
	×	✓	✓	69.61	0.07	0.44
	✓	×	✓	71.08	0.17	0.58
	✓	✓	✓	70.05	0.08	0.45
VisualBERT	×	×	×	64.92	0.14	0.50
	×	✓	×	63.98	0.07	0.40
	×	✓	✓	66.18	0.06	0.40
	✓	×	✓	65.67	0.07	0.43
	✓	✓	✓	65.26	0.03	0.36
ViLBERT	×	×	×	66.98	0.19	0.57
	×	✓	×	66.44	0.12	0.47
	×	✓	✓	68.02	0.10	0.46
	✓	×	✓	67.49	0.11	0.48
	✓	✓	✓	66.90	0.05	0.39

Table 7: Effectiveness of each component in terms of improving the calibrating performance

VQA Model	Ensemble Size	Acc.	ECE	OC	AUC
ViLBERT	2	66.38	0.06	0.41	12.65
	3	66.90	0.05	0.39	12.22
	4	67.17	0.04	0.38	11.94
	5	67.38	0.04	0.39	11.78
VisualBERT	2	64.69	0.03	0.35	13.89
	3	65.26	0.03	0.36	13.30
	4	65.90	0.02	0.34	12.66
	5	65.60	0.03	0.33	12.92

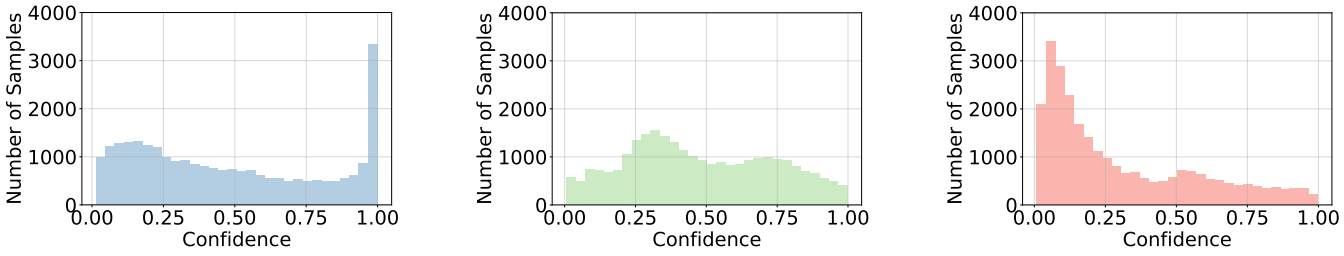
Table 8: Ablation study on the effect of number of sub-networks within ensemble on the performance.

and training it from scratch not only adversely affects the learning capability but also does not provide the same calibration benefits as LRF. We further extended this experiment to the entire GLEN framework.

Similarly, We designed individual GFL-based VQA models with bottleneck layers, trained them with the GFL loss from scratch, and finally ensembled them together. This model achieved an accuracy of 65.50 and an ECE of 0.09. In comparison, the ensemble of GFL-based standard VQA models (with original layers) achieves an accuracy of 66.30 and an ECE of 0.10. On the other hand, GLEN (i.e., the ensemble of GFL-based VQA models followed by LRF) achieves an accuracy of 66.15 and an ECE of 0.06. This study demonstrate that simply designing models with bottleneck layers and training from scratch does not provide the same calibration benefits as applying LRF post-training. The LRF step is crucial for improving the calibration of VQA models while preserving their learning capabilities. Directly using the bottleneck design may limit the model from learning important features that hurts its generalization capability.

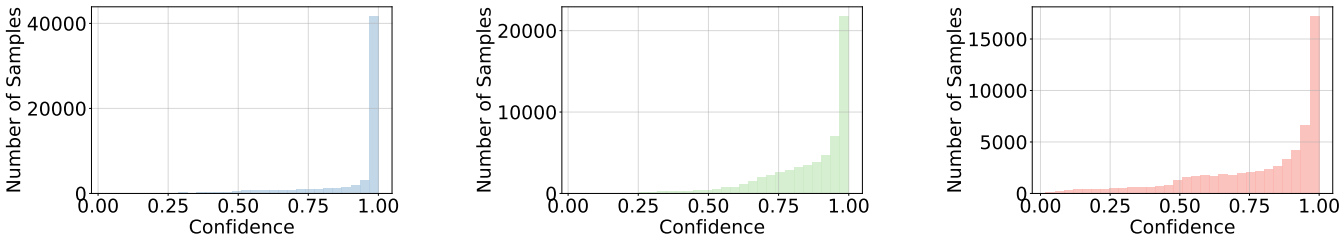
F.6 Qualitative Analysis

We conduct an additional qualitative analysis to showcase the effectiveness of the proposed `GLEN` on the VQA-v2 dataset using the ViLBERT backbone. figs. 12 and 13 show the histogram plots of the confidence scores for the correctly and incorrectly classified samples using `GLEN`, Baseline, and Selector. In case of incorrect data samples, the proposed `GLEN` produces a small confidence score for most of the samples whereas in the case of Baseline and Selector, the confidence scores for those samples remain high. In contrast, in the case of the correct samples, `GLEN` produces a high confidence score. This justifies that `GLEN` is better calibrated, which produces low confidence in case of incorrect samples. Additionally, fig. 14 depicts the confidence histograms of the three methods for partially correct answers, particularly those having an accuracy of 0.3. Similar to incorrect answers, as fig. 14a shows, the baseline model is highly overconfident, as the majority of confidence scores are concentrated



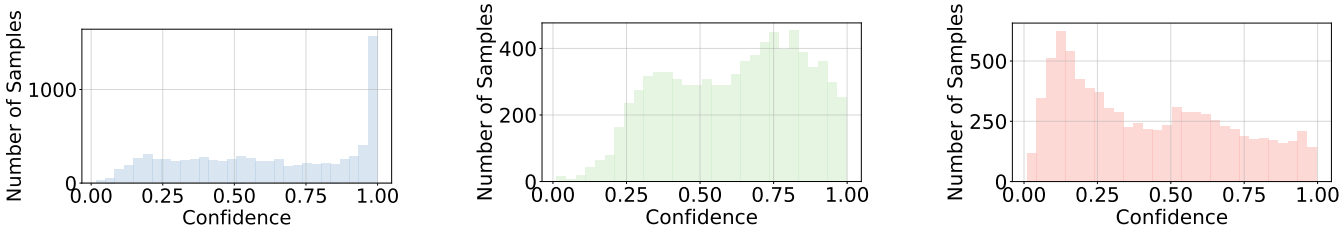
(a) Baseline (b) Selector (c) GLEN

Figure 12: Confidence histograms of wrongly answered questions by a) Baseline, b) Selector, c) GLEN



(a) Baseline (b) Selector (c) GLEN

Figure 13: Confidence histograms of correctly answered questions by a) Baseline, b) Selector, c) GLEN



(a) Baseline (b) Selector (c) GLEN

Figure 14: Confidence histograms of questions answered partially correctly (with accuracy 0.3) by a) Baseline, b) Selector, c) GLEN

Model	LRF Approach	Acc.	ECE	OC	AUC
CLIP-ViL	Uniform LRF	70.05	0.08	0.45	10.46
	Non-Uniform LRF	70.23	0.07	0.44	10.27
ViLBERT	Uniform LRF	66.90	0.05	0.39	12.22
	Non-Uniform LRF	67.11	0.05	0.40	12.10

Table 9: Uniform vs. non-uniform factorization rank

Model	Post-train. LRF		Bottleneck Layer	
	Acc↑	ECE↓	Acc↑	ECE↓
Standard	**65.4**	**0.10**	64.4	0.14
GFL Ensemble	**66.2**	**0.06**	65.3	0.09

Table 10: Post-training LRF vs. Bottleneck model

around 1. Selector alleviates the overconfidence issue slightly, by shifting the confidence scores to the lowers side. However, the majority still lies in the high confidence bins. On the contrary, as shown in fig. 14c, GLEN assigns a much lower confidence score, compared to Selector, to such answers with a low accuracy.

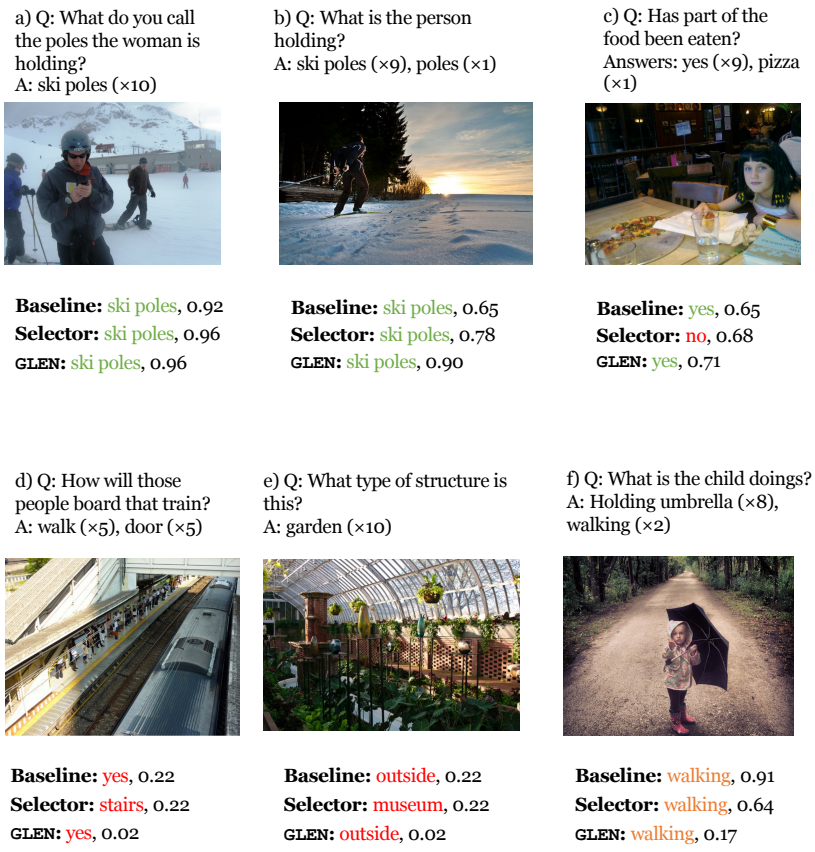


Figure 15: Illustrative examples showing (a-c) correct, (d-e) incorrect, (f) partially correct answers by GLEN and other baselines

fig. 15 show the examples with the respective prediction and the associated confidence score for different methods. In the case of the correct samples, GLEN remains confident whereas in the case of the incorrect samples, it lowers the confidence. Furthermore, Baseline and Selector remain confident even in case of incorrect samples. Additionally, in the partially correct

case, where Baseline is overconfident, GLEN exhibits a lower confidence, as compared to Selector. Those examples help to further justify the better calibration ability of our proposed technique compared to competitive baselines.

F.7 Experimental Results on VizWiz dataset: A challenging VQA dataset

In this study, we conducted experiments using the VizWiz dataset (Gurari et al. 2018), a visual question answering (VQA) dataset that is uniquely suited to real-world applications due to its collection by blind and visually impaired individuals. The dataset consists of images captured using phones, accompanied by questions (spoken) about the images. Given the nature of the data collection process, the images often exhibit characteristics such as blurriness, lack of clarity, or incomplete framing of the objects of interest, resulting in a substantial portion of unanswerable questions. Each image/question pair in the dataset is provided with ten potential answers.

The most recent version of the VizWiz dataset, released in 2020, contains 20,523 image/question pairs in the training set, 4,319 pairs in the validation set, and 8,000 pairs in the test set. However, as the answers for the test set are not publicly available, and accurate calibration of models necessitates access to instance-level accuracies, we opted to re-split the dataset. Specifically, we designated the provided validation set as our new test set and created a new validation set by randomly selecting 2,000 image/question pairs from the original training set. Because each question is associated with a unique image, the random splitting was straightforward.

Furthermore, to accommodate ad-hoc calibration methods such as Selector, Vector Scaling, and temperature scaling—which require an additional validation set for training the calibration component—we extracted a further subset of 1,000 instances from the training set to serve as a dedicated calibration validation set. This structured approach ensures that the dataset is appropriately partitioned for both model evaluation and calibration.

Table 11 presents the performance comparisons on ViLBERT, VisualBERT and BEiT-3 models. Our GLEN consistently achieves lowest ECE, and OC compared to all with slightly higher accuracy than other methods.

Model		Acc↑	ECE↓	OC↓	AUC↓	C@1↑	C@5↑	C@10↑	C@20↑
VisualBERT	**Baseline**	49.70	0.39	0.69	30.36	**0.51**	0.80	3.40	25.43
	VectorScale	49.95	0.15	0.43	30.57	0.07	0.41	1.43	23.05
	Select	49.70	0.15	0.46	**28.61**	0.13	**1.34**	**7.82**	**31.85**
	GLEN	**51.02**	**0.12**	**0.42**	28.96	0.41	1.13	6.43	28.05
ViLBERT	**Baseline**	50.46	0.31	0.59	28.67	0.57	2.36	8.45	26.59
	VectorScale	50.53	0.11	0.38	28.37	0.09	0.09	0.32	33.28
	Select	50.46	0.17	0.46	27.75	**1.36**	**3.18**	**11.47**	33.75
	GLEN	**50.87**	**0.09**	**0.35**	**26.04**	0.53	2.39	10.75	**38.00**
BEiT-3$_{\text{base}}$	**Baseline**	**68.51**	0.08	0.51	**12.57**	6.32	25.82	**47.01**	**75.76**
	VectorScale	68.44	0.10	0.55	12.78	5.46	25.71	46.20	74.84
	Select	68.45	0.07	0.55	12.94	**10.85**	24.52	43.70	74.97
	GLEN	68.50	**0.05**	**0.46**	12.85	10.20	**26.88**	43.71	74.36

Table 11: Performance comparison on a vizwiz dataset.

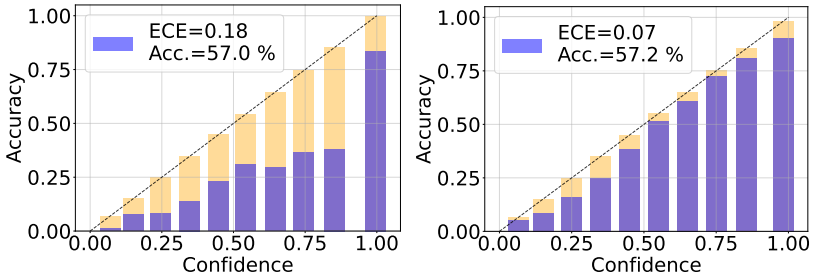
F.8 A Critical Domain Evaluation: Medical-VQA

To demonstrate the effectiveness of our proposed technique in a critical domain where reliability is crucial, we conducted an evaluation within the medical Visual Question Answering (VQA) domain, specifically on the PathVQA dataset (He et al. 2020). The PathVQA dataset, comprising of $32,799$ questions and $4,998$ pathology images, presents a unique challenges in the medical VQA domain. The inherent scarcity of annotated question-image pairs in medical domain, due to the difficulty of obtaining annotated question image pairs, results in significantly smaller datasets when compared to standard VQA datasets like VQA-v2 (Goyal et al. 2017). This limitation, makes medical VQA models prone to issues such as overfitting, poor calibration, and overconfidence, which are particularly concerning given the critical nature of medical applications. The importance of model calibration is underscored in these settings to ensure reliability and trustworthiness in predictions.

Evaluation compares GLEN against a baseline model based on a Bilinear Attention Network (BAN) (Kim, Jun, and Zhang 2018), in terms of calibration, accuracy and abstention performances. The baseline model, exhibits a poor calibration, with an Expected Calibration Error (ECE) of 0.19 as shown in fig. 16a, and poor abstention performances specifically on the C@1, C@5, and C@10 metrics. In contrast, GLEN significantly enhances model calibration, reducing the ECE to 0.07, with a slight improvement of 0.2% in VQA accuracy relative to the baseline as in Figure 16b. Moreover, the abstention performances C@1, C@5, C@10 are substantially improved. Notably, the evaluation reveals that while ensembling baseline models marginally increases accuracy, it fails to improve calibration errors, and abstention performances. These findings, detailed in Figure 12, underscore the superiority of GLEN in both improving calibration and maintaining competitive accuracy in the medical VQA domain, thereby addressing critical limitations of existing approaches.

Model	Acc↑	ECE↓	AUC↓	C@1↑	C@5↑	C@10↑	C@20↑
Baseline (He et al. 2020)	57.0	0.19	24.0	0	0	0	64.9
Baseline Ensembles	57.7	0.18	23.5	0	0	0	65.7
GLEN	57.2	**0.07**	**17.4**	**2.5**	**6.8**	**41.0**	**65.1**

Table 12: Performance comparison on a Medical VQA dataset.



(a) Baseline (b) GLEN

Figure 16: ECE plot of a) Baseline, b) GLEN on a medical VQA dataset: PathVQA (He et al. 2020).

F.9 Effectiveness of GLEN on Large Foundation Models

Due to constraints on the page limits, in this section we provide further details, and the full comparison using the BEiT-3 model including the comparison with selector, on the VQA-v2 dataset in the Table 13. In this experiment we used a BEiT-3 *base* model with 228M parameters. All models are fine-tuned on the VQA-v2 dataset (Goyal et al. 2017) for 10 epochs. In GLEN, the ensembled models were fine-tuned with the Generalized Focal Loss (GFL). As illustrated in table 13 GLEN achieves ECE improvement from 0.09 to 0.02, maintaining a similar VQA accuracy of 74%. Additionally, GLEN improves the selective prediction performances over the baseline, demonstrating a comparable performance to Selector. This study demonstrates our method's compatibility and effectiveness with current state-of-the-art models, demonstrating its relevance and adaptability to ongoing advancements in the field.

Parameter Efficient Ensembling in Large State-of-the-art Models. Ensembling large SOTA models might pose practical challenges due to their huge parameter size. To mitigate this, we propose using parameter-efficient fine-tuning techniques *e.g.* LoRA (Hu et al. 2021) for fine-tuning adaptation layers within each ensemble, while keeping the base pretrained large model frozen. This approach significantly cuts computational and storage needs of fine-tuning multiple subnetworks for GLEN, and ensembling at inference time.

To empirically validated this strategy we employed LoRA for the fine-tuning of BEiT-3 models. For GLEN the LoRA adaptation layers are fine-tuned with the GFL loss. Each model is fine-tuned for 20 epochs, with a fixed adaptation layer rank of 16. As indicated by table 14, the results demonstrate that low-rank fine-tuning of the BEiT-3 model decreased the ECE from 0.09 to 0.05, when compared to the fully fine-tuned BEiT-3 model. This improvement in calibration could be attributed to the reduced overfitting due to the low-rank finetuning. Furthermore, GLEN achieves a further reduction in the ECE to 0.03 with only a marginal compromise in accuracy. It is important to note, that the fine-tuning process via LoRA may not be fully optimized regarding the chosen hyperparameters and the number of fine-tuning epochs, as evidenced by the observed reduction in accuracies.

F.10 Experiments Reproducibility

In this section the hyperparameters used for the experiments are presented. Specifically, we used two key hyperparameters (λ) (in eq. (4)) and LRF ratio. As eq. (4) involves inequality constraint, which incurs a higher computational overhead, in our actual

Model		Acc.↑	ECE↓	OC↓	AUC↓	C@1↑	C@5↑	C@10↑	C@20↑
BEiT-3	Baseline	74.68	0.09	0.55	7.78	14.86	47.62	66.89	91.41
	VectorScale	74.51	0.08	0.54	7.81	18.01	48.16	57.40	91.12
	Selector	74.61	0.06	0.53	7.56	**20.01**	**49.17**	**68.46**	90.93
	GLEN	**74.95**	**0.02**	**0.43**	**7.43**	17.93	48.66	68.36	**91.67**

Table 13: Performance comparison on VQA-v2 test split, for BEiT-v3 foundation model.

Model	Acc↑	ECE↓
BEiT-3$_{\text{base}}$+LoRA	71.6	0.05
GLEN+LoRA	71.1	**0.03**

Table 14: Efficient Ensembling of large models in GLEN by LoRA.

optimization process, we consider the equivalent regularized version of the GFL as follow (Sapkota and Yu 2023):

$$\mathcal{L}(\Theta)^{GFL} = \max_{\mathbf{w},\mathbf{w}^T\mathbb{I}=1} \sum_{n=1}^{N} w_n l(\mathbf{x}_n, \Theta) - \beta D_f \left(\mathbf{p}||\frac{\mathbb{I}}{N}\right) \tag{23}$$

Solving this inequality easily leads to the following closed form solution [1]:

$$\mathcal{L}(\Theta)^{GFL} = \sum_{n=1}^{N} w_n^* l(\mathbf{x}_n, \Theta) \tag{24}$$

Where, w_n^* is given as

$$w_n^* = \frac{\exp(\frac{l(\mathbf{x}_n,\Theta)}{\beta})}{\sum_{j=1}^{N} \exp(\frac{l(\mathbf{x}_j,\Theta)}{\beta})} \tag{25}$$

Now, we have new hyperparameter β based on the equivalent regularized version of GFL. As our model is based on the ensemble model, we use different β for base learners and are indicated as β_1, β_2 and β_3 respectively. Table 15 demonstrates hyperparameters for reproducibility. We average the results over 4 runs.

	VQA Model	β_1	β_2	β_3	**LRF ratio**
VQA-v2	LXMERT	10	20	1000	0.3
	Pythia	8	100	1000	0.9
	ViLBERT	8	20	100	0.7
	VisualBERT	10	20	100	0.9
	CLIP-ViL	20	100	1000	0.8
	BEiT-3	8	200	500	0.7
	VQA Model	β_1	β_2	β_3	**LRF ratio**
Vizwiz	ViLBERT	2	6	20	0.3
	VisualBERT	0.1	1	10	0.8
	BEiT-3	2	15	1000	0.9

Table 15: Final hyperparameters of experiments presented in the paper, for VQA-v2, and Vizwiz datasets.

G Broader Impact

Improving the calibration of VQA systems, is an important step towards enhancing reliability and trustworthiness of VQA systems, which enhances user confidence, facilitating the wider adoption of these technologies in various real-world domains. Firstly, by mitigating the risks associated with overconfidence and poor calibration in VQA models, our research paves the way for more dependable systems that users can trust to make informed decisions. In critical domains such as medical diagnostics, where VQA models assist healthcare professionals by providing insights from medical imagery, the accuracy and reliability of such systems can significantly impact patient outcomes. Improving model calibration ensures that the confidence levels associated with answers are more reflective of actual model performance, thereby reducing the likelihood of erroneous decisions based on misguided trust in the system's capabilities. Secondly, in applications where user interacts with AI frequently, such as educational tools, the trust established through answers and the corresponding calibrated confidences can enhance user engagement and satisfaction. Users are more likely to rely on and return to a system that consistently demonstrates a high level of accuracy and acknowledges its limitations by abstaining from answering when unsure. Moreover, our work has the potential to inspire future research in vision-language modeling, encouraging the exploration of novel methods for improving the calibration and reliability of models across various applications. This can lead to further contributions towards development of AI technologies that are not only more accurate but also more aligned with human values and needs.

H Limitations and Future Works

The GLEN approach enhances the calibration of Visual Question Answering (VQA) models, thereby increasing the reliability of estimated confidence scores associated with the VQA's predicted answers, specifically reducing the over-confident predictions. To assess the reliability of the system, a selective prediction strategy was employed. While this strategy effectively reduces the risk of providing incorrect answers, it introduces a notable limitation as the questions remain unanswered, leading to a poor user-experience.

A promising direction for future research is the integration of specialized models for additional answer verification or to handle questions where the primary VQA model shows low confidence. Specifically, leveraging Large Language Models (LLMs) for further validation of answers, based on the calibrated confidence levels, presents an intriguing extension. LLMs, with their extensive knowledge base and advanced reasoning capabilities, could significantly contribute to verifying answers and filling gaps where the primary VQA model is not confident. On the otherhand, directly utilizing LLMs, such as GPT-4, for answering all questions presents practical limitations primarily due to the computational and financial costs associated with fine-tuning, and inference. Hence, the strategic integration of LLMs for specific tasks, such as verifying answers where the VQA model's confidence is low, emerges as a more feasible and cost-effective approach. This method ensures that LLMs are utilized effectively, by applying their advanced reasoning capabilities where they add the most value without incurring prohibitive computational and financial costs. By focusing on this hybrid approach, future research can aim to not only enhance the accuracy and reliability of VQA systems but also maintain a balance between performance, user engagement, and operational costs. The discussed evolution of VQA systems aims at ensuring a more reliable and user-centric approach, encouraging broader acceptance and reliance on these systems for information and assistance.

References

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.

Ben-Younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2612–2620.

Cao, R.; and Li, Z. 2023. Overcoming Language Priors for Visual Question Answering via Loss Rebalancing Label and Global Context. In *Uncertainty in Artificial Intelligence*, 249–259. PMLR.

Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6639–6648.

Gao, Y.; Beijbom, O.; Zhang, N.; and Darrell, T. 2016. Compact bilinear pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 317–326.

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.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. PathVQA: 30000+ Questions for Medical Visual Question Answering. *arXiv preprint arXiv:2003.10286*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. *Advances in neural information processing systems*, 31.

Kim, J.-H.; On, K.-W.; Lim, W.; Kim, J.; Ha, J.-W.; and Zhang, B.-T. 2016. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

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.

Lu, P.; Li, H.; Zhang, W.; Wang, J.; and Wang, X. 2018. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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.

Sapkota, H.; and Yu, Q. 2023. Adaptive Robust Evidential Optimization For Open Set Detection from Imbalanced Data. In *The Eleventh International Conference on Learning Representations*.

Schwartz, I.; Schwing, A.; and Hazan, T. 2017. High-order attention models for visual question answering. *Advances in Neural Information Processing Systems*, 30.

Shao, Z.; Yu, Z.; Wang, M.; and Yu, J. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14974–14983.

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*.

Singh, A.; Goswami, V.; Natarajan, V.; Jiang, Y.; Chen, X.; Shah, M.; Rohrbach, M.; Batra, D.; and Parikh, D. 2020. Mmf: A multimodal framework for vision and language research. *MMF: A multimodal framework for vision and language research*.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *International Conference on Learning Representations*.

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.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6281–6290.