Variational Visual Question Answering for Uncertainty-Aware Selective Prediction

Anonymous authors
Paper under double-blind review

Abstract

Despite remarkable progress in recent years, vision language models (VLMs) remain prone to overconfidence and hallucinations on tasks such as Visual Question Answering (VQA) and Visual Reasoning. Bayesian methods can potentially improve reliability by helping models *selectively predict*, that is, models respond only when they are sufficiently confident. Unfortunately, Bayesian methods are often assumed to be costly and ineffective for large models, and there exists little evidence to show otherwise for multimodal applications. Here, we show the effectiveness and competitive edge of variational Bayes for selective prediction in VQA for the first time. We build on recent advances in variational methods for deep learning and propose an extension called "Variational VQA". This method improves calibration and yields significant gains for selective prediction on VQA and Visual Reasoning, particularly when the error tolerance is low ($\leq 1\%$). Often, just one posterior sample can yield more reliable answers than those obtained by models trained with AdamW. In addition, we propose a new risk-averse selector that outperforms standard sample averaging by considering the variance of predictions. Overall, we present compelling evidence that variational learning is a viable option to make large VLMs safer and more trustworthy.

1 Introduction

Advances in Vision Language models (VLMs) (Wang et al., 2023; 2024; Li et al., 2024) have led to substantial gains on classical Visual Question Answering benchmarks (Antol et al., 2015; Goyal et al., 2016), with performance now approaching or surpassing human levels. However, even strong VQA models are miscalibrated, prone to hallucinations, and often confidently guess answers instead of expressing uncertainty (*cf.* Fig. 1). In a nutshell, these models "don't know what they know" - a shortcoming which hinders their deployment in safety-critical domains such as medical diagnosis or assistance for the visually impaired. These issues become even more pronounced in novel situations, such as adversarial (Sheng et al., 2021) or unanswerable (Bigham et al., 2010) inputs, which are common in the real world.

Abstentions are formalized in the selective prediction framework Chow (1957). Although selective prediction has recently received attention in the context of hallucinations (Kalai et al., 2025), the literature on multimodal models remains sparse. Previous approaches have relied on additional model components: Whitehead et al. (2022) train a lightweight head on top of the frozen VLM backbone, while Srinivasan et al. (2024) use external vision tools and an additional language model to quantify uncertainty. In both cases, the underlying predictive model remains unreliable, and while additional training increases overhead, external tools add vulnerabilities that require careful design.

Bayesian models (Blundell et al., 2015) can potentially address the unreliability of VLMs without requiring additional components or tools. In particular, the uncertainty in the learned posterior distribution over model parameters can be used to help the model make a prediction only when it is sufficiently confident. This theory remains untested though, as for a long time, Bayesian approaches have been ineffective for large transformer architectures (Khan et al., 2018). However, recent progress in variational learning (Shen

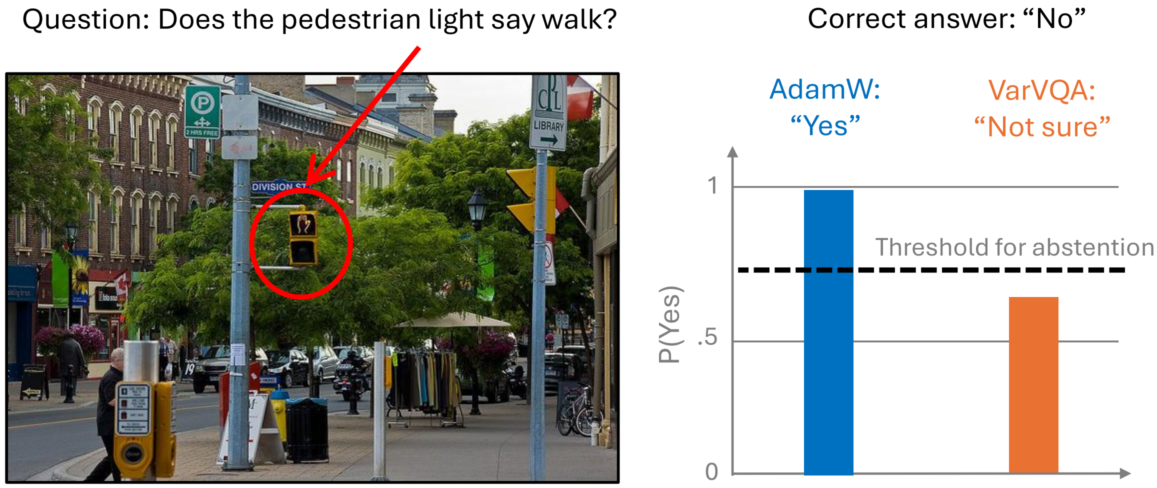


Figure 1: Despite recent performance gains, VLMs trained with popular optimizers like AdamW do not know when they are wrong. Our Variational VQA approach uses learned parameter variances to enable models to abstain when uncertain. The example is from BEiT-3 (Wang et al., 2023), which achieves near-human accuracy on VQAv2 (Goyal et al., 2016).

et al., 2024) has enabled effective training of unimodal models such as GPT-2 (Radford et al., 2019) with no significant training overhead compared to the common AdamW optimizer (Loshchilov and Hutter, 2019).

In this work, we are the first to extend the recent IVON (Shen et al., 2024) method to the multimodal domain and comprehensively demonstrate its effectiveness for selective prediction. Models trained with IVON learn parameter variances, which we use for uncertainty estimation in VQA. Our contributions are as follows.

1. We introduce Variational VQA (VarVQA) as a framework for intrinsic uncertainty estimation in multimodal models. We are the first to apply IVON to VLMs, demonstrating that variational training is effective for large multimodal architectures without sacrificing accuracy or incurring significant training overhead.
2. We demonstrate improved uncertainty estimation across multiple dimensions: Better calibration, enhanced selective prediction (with particularly large gains at low error tolerances), and increased robustness under distribution shift.
3. We establish superior sample efficiency compared to Monte Carlo (MC) Dropout, showing that Variational VQA provides better reliability when the invested compute is equal.
4. We propose a new risk-averse selector function that leverages output variance, yielding consistent improvements in *high-stakes* selective prediction where errors are particularly costly.

2 Related Work

Visual Question Answering. Visual Question Answering (VQA) is a popular multimodal task that requires a model to understand two modalities and their interaction to predict answers, which makes uncertainty estimation challenging. As recent models (Li et al., 2023; Wang et al., 2023; 2024) have achieved near-human level performance on standard VQA datasets like VQAv2 (Goyal et al., 2016), the community has moved to newer VQA benchmarks that test more diverse capabilities, like MMBench (Liu et al., 2024) and MME (Fu et al., 2024). However, even models that reach near-human accuracy on VQAv2 still perform poorly when evaluated in terms of selective prediction (Dancette et al., 2023). In this work, we show, for the first time, the effectiveness of Bayesian methods (Shen et al., 2024) to address abstentions in large VLMs.

Selective Prediction. In the selective prediction framework (Chow, 1957; El-Yaniv and Wiener, 2010), a selection function or "selector" takes the role of assigning a confidence to a given model answer. The decision whether *a)* a model's response is accepted or *b)* it abstains (*i.e.* it says "I don't know") is then made using this confidence and an abstention threshold. If the confidence is below the threshold, the model abstains, but otherwise the prediction is accepted. Typically, the highest answer likelihood (Geifman and El-Yaniv, 2017) or the predictive entropy are used as a selection function. Most prior work on selective prediction can be classified into external and integrated approaches. In external setups, a selector is built on top of the frozen predictive model, *e.g.* in the form of a trainable model head (Whitehead et al., 2022; Mielke et al., 2022; Dancette et al., 2023; Mushtaq et al., 2025), LoRA parameters (Chen et al., 2023) or vision tools (Srinivasan et al., 2024). In integrated setups, predictor and selector have at least one combined training phase. Integrated selectors take different forms as well, such as a model head (Geifman and El-Yaniv, 2019) or a dedicated abstention class (Ziyin et al., 2019). However, if model and selector are trained together, instabilities often ensue, which need special treatment (Geifman and El-Yaniv, 2019). Bayesian approaches have not been considered for selective prediction so far, with the exception of concurrent work by (Daheim et al., 2025), which has explored IVON for generative language modeling, but not for multimodal tasks. In contrast to prior work on selective prediction in VQA, our objective is to *directly* improve the reliability of model confidence estimates without additional parameters, training phases, or tools. In other words, we train VLMs where reliability is "baked-in" by design, not added as an afterthought.

Calibration. Calibration represents a different angle on uncertainty estimation, namely the alignment of a model's predictive confidence with its accuracy. In other words, when a model expresses $x\%$ confidence in an answer, it should be correct $x\%$ of the time. The difference to selective prediction becomes clear when considering a model that is right on $x\%$ of examples and always expresses the same confidence of $x\%$. Although such a model is perfectly calibrated, it cannot distinguish its correct and incorrect outputs and thus cannot help with the task of deciding when to abstain. Prior work has found that large neural networks often exhibit overconfidence, particularly in OOD settings (Snoek et al., 2019). In unimodal classification tasks, temperature (Guo et al., 2017) and Platt (vector) Platt et al. (1999) scaling are effective at improving calibration. Ensembling (Lakshminarayanan et al., 2017) typically yields even better results, but requires prohibitive resources to train N models. New ideas, such as prompting the model to express a verbalized confidence have been mostly ineffective for VLMs (Xuan et al., 2025). We show that Variational VQA yields well-calibrated VLMs, achieving a lower Expected Calibraion Error (ECE) than vector scaling, while matching other sampling methods like Monte-Carlo Dropout (Gal and Ghahramani, 2016). In general, we argue that for a VLM to be reliable, it should *a)* be calibrated and *b)* know when to abstain - both of these aspects are much improved with Variational VQA compared to standard AdamW training.

Variational Learning. Variational Learning provides a principled approach to estimate uncertainty by learning probability distributions (often Gaussians) over network weights. In the early 2010s, promising results were achieved by variational methods that directly optimize parameter means and variances through standard deep learning techniques such as SGD (Graves, 2011; Blundell et al., 2015). However, these approaches could not keep up with the growth in scale of network architectures in subsequent years (Trippe and Turner, 2018; Foong et al., 2020; Coker et al., 2022). Recent works employing *natural gradients* (Khan et al., 2018; Osawa et al., 2019) build an estimate of the Hessian matrix through an Adam-like update. IVON (Shen et al., 2024) further develops those and can obtain comparable accuracy and better uncertainty estimates than AdamW at nearly identical training cost. We use IVON because it offers several advantages compared to other Bayesian baselines. Unlike the Laplace approximation (MacKay, 1992; Daxberger et al., 2021), it does not require an additional pass through the data to compute the Hessian. Neither does it require additional training like Stochastic Weight Averaging (SWA) (Izmailov et al., 2018). Compared to MC Dropout (Gal and Ghahramani, 2016), the advantage is the availability of a fixed posterior form that can be more easily used for downstream tasks. For instance, the method is easily amenable to ensembling (Lakshminarayanan et al., 2017), which can further improve performance (Daheim et al., 2025).

We offer new insights compared to previous IVON works (Shen et al., 2024; Cong et al., 2025; Daheim et al., 2025), by showing its effectiveness in training multimodal models and for selective prediction. We further propose a new selection function that uses the output variance, which was never utilized in prior work.

3 Variational Learning and Selective Prediction

We explain the variational learning paradigm in Sec. 3.1, briefly describe the IVON optimizer (Sec. 3.2), and formalize selective prediction in VQA (Sec. 3.3).

3.1 Variational Learning

Deep learning methods estimate network weights θ by minimizing *empirical risk* $\ell(\theta) = \frac{1}{M}\sum_{k=1}^{M}\ell_k(\theta)$, where M is the size of the training set and $\ell_k(\theta)$ the loss for example k. In contrast, *variational* learning methods aim to estimate a distribution $q(\theta)$ over network parameters by minimizing

$$\mathcal{L}(q(\theta)) = \lambda\mathbb{E}_{q(\theta)}\left[\ell(\theta)\right] + \mathbb{D}_{\mathrm{KL}}(q(\theta) \parallel p(\theta)). \tag{1}$$

Here, $\mathbb{D}_{\mathrm{KL}}$ is the Kullback-Leibler divergence, $\lambda \approx M$ a scaling parameter and $p(\theta)$ the prior distribution over weights. To keep computational costs manageable, the distribution over weights is often chosen to be a diagonal covariance Gaussian, that is, we set $q(\theta) = \mathcal{N}(\theta \mid m, \mathrm{diag}(V))$, where m and V are the parameter mean and parameter variance vectors, respectively. The loss $\mathcal{L}(q(\theta)) = \mathcal{L}(m, V)$ is typically approximated through MC sampling of the model parameters.

3.2 IVON

The IVON optimizer (Shen et al., 2024) uses an Adam-like (Kingma, 2014) update for the parameter means m and variances V, where the Hessian estimate h takes the role of the momentum. Essentially, m is updated using gradients scaled by h. A notable difference to Adam and its variants is the absence of the square root over the momentum term $(h + \delta)$. The updates made in every training step are detailed below.

$$\hat{h} \leftarrow \frac{\hat{g}(\theta - m)}{V}, \tag{2}$$

$$m \leftarrow m - \alpha \cdot \frac{g + \delta m}{h + \delta}, \tag{3}$$

$$V \leftarrow \frac{1}{\lambda \cdot (h + \delta)}. \tag{4}$$

Here, α is the learning rate and δ the weight decay. IVON also uses Adam-like momentum for the gradients g and the Hessian h: In Equation (2), the variables $\hat{h}$ and $\hat{g}$ refer to estimates in the current step, while g and h in Equations (3) and (4) represent the smoothed average. To obtain reasonable parameter uncertainties V, the Hessian needs to be initialized, typically by a constant h_0. For more details, we refer to the original paper by Shen et al. (2024).

3.3 Selective Prediction in VQA

In VQA, the model learns a function $f : \mathcal{I} \times \mathcal{Q} \to \mathcal{A}$ to predict an answer $a \in \mathcal{A}$, given a multimodal input $x = (i, a)$ consisting of an image $i \in \mathcal{I}$ and a question $q \in \mathcal{Q}$. In the selective prediction framework, the model output space is augmented by an *abstain* output $\emptyset$. This transforms the predictive model f into a selective model h, incorporating both f and a selector g. The answer $f(x)$ is accepted if $g(x)$ is above the abstention threshold γ, and rejected otherwise. We follow the notation of Whitehead et al. (2022):

$$h(x) = (f, g)(x) = \begin{cases} f(x) & \text{if } g(x) \geq \gamma, \\ \emptyset & \text{if } g(x) < \gamma. \end{cases} \tag{5}$$

A high threshold γ corresponds to a conservative case, in which the model answers only the questions on which it is most confident. Lowering γ reduces the number of abstentions, but increases the error rate. In practice, γ is set according to a pre-specified cost of error or desired error rate, see Section 5.2.

4 Variational VQA

In essence, our Variational VQA approach uses the IVON optimizer to train large VLMs and evaluates the reliability of its output confidences in comparison to baselines like AdamW and MC Droput. In Section 4.1, we describe how model confidences are obtained, in Section 4.2 we describe the baseline selectors, and in Section 4.3 we present our new risk-averse selector.

4.1 Inference and Model Confidence

At inference, variational methods typically make use of the learned posterior distribution through Monte-Carlo (MC) sampling. However, if computing efficiency is imperative, one can ignore the variances ($V = 0$) and use only the mean parameters m for inference (Shen et al., 2024). This requires only one forward pass. We refer to this approach as 'VarVQA mean'. For an input x, the output distribution vector is $\tilde{p}(x)$.

VarVQA performs sampling, *i.e.* we do $n \in \mathbb{N}$ MC samples of the model parameters and obtain output distribution vectors $\tilde{p}_n$, where $K \in \mathbb{N}$ is the number of classes. These are aggregated to obtain an *output* mean vector $\tilde{\mu}$ and an *output* variance vector $\tilde{\sigma}$ for every input x:

$$\tilde{\mu}(x) = \frac{1}{N} \sum_{n=1}^{N} \tilde{p}_n(x) \tag{6}$$

$$\tilde{\sigma}^2(x) = \frac{1}{N-1} \sum_{n=1}^{N} \left(\tilde{\mu}(x) - \tilde{p}_n(x)\right)^2 \tag{7}$$

4.2 Baseline selector functions

We start by explaining the baseline selector for deterministic methods (AdamW, VarVQA mean). We employ the widely used MaxProb (Geifman and El-Yaniv, 2017), which uses the highest answer likelihood. Let $\tilde{p}(x)$ be the model output; then the MaxProb selector is defined as $g_{\mathrm{MP}}(x) = p^{\star}(x) = \max_k \tilde{p}^k(x)$ (here, k enumerates the output classes). We find that MaxProb consistently outperforms predictive entropy and related functions.

In case of multiple MC samples, the default method in the field of uncertainty estimation is predictive averaging (Gal and Ghahramani, 2016). In essence, predictive averaging is an application of MaxProb on the mean output distribution, *i.e.* $g^{\mu}_{\mathrm{MP}}(x) = \mu^{*}(x) = \max_k \tilde{\mu}^k(x)$ (*cf.* Eq. (6)).

4.3 A new risk-averse selector

In this work, particularly for the context of selective prediction, we propose to go *Beyond Predictive Averaging* (BPA) by also employing the output variances (*cf.* Eq. (7)). This is done in a risk-averse (Pratt, 1978) manner, by penalizing high-variance predictions. While Pratt (1978) subtracts the variance (with a prefactor), we found the standard deviation to work best:

$$g_{\mathrm{BPA}}(x) = \mu^{*}(x) - \sigma^{*}(x) \tag{8}$$

Here, σ^* is the variance of the highest-likelihood class, *i.e.* the risk-averse selector does not change the prediction, only the confidence. All our selective prediction results with VarVQA use g_{BPA} by default. In Section 5.6, we provide an ablation against predictive averaging. When it comes to calibration, VarVQA uses predictive averaging, as the subtraction of σ leads to systematic underconfidence[1]. When using MC Dropout with AdamW, we found no systematic benefits of g_{BPA}. We speculate that this is because the posterior was not actively learned. Thus, we use only g^{μ}_{MP} for Dropout. The selectors used for each method are visually summarized in Figure 2.

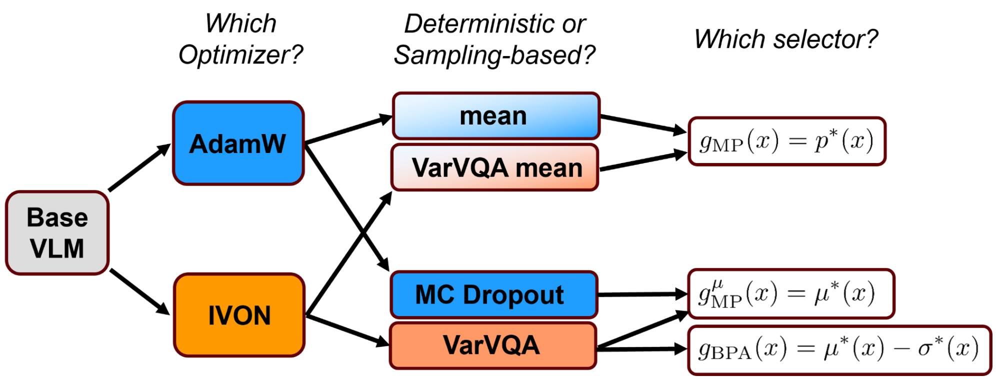


Figure 2: Overview of the methods we experiment with and their selectors. Variational VQA employs g^{μ}_{MP} for calibration and g_{BPA} for selective prediction.

5 Experiments

We describe our experimental setup, models and datasets in Section 5.1 and the evaluation metrics in Section 5.2. Our results show that Variational VQA is effective for multimodal models (Sec. 5.3), more sample-efficient than MC Dropout (Sec. 5.4), and more robust to OOD data than AdamW-trained models (Sec. 5.5). Moreover, our novel selector g_{BPA} outperforms posterior predictive averaging on high-stakes selective prediction (*cf.* Sec. 5.2) across multiple models and tasks (Sec. 5.6).

5.1 Experimental Setup

We explore the effectiveness of Variational VQA on two large VLMs: ViLT (Kim et al., 2021) and BEiT-3 (Wang et al., 2023). BEiT-3 s near-SOTA[2] on VQAv2, but still small enough for full fine-tuning. Both ViLT and BEiT-3 treat VQA as a classification task to 3129 answers, which is standard practice (Anderson et al., 2018). In terms of multimodal tasks, we explore VQA (fine-tuning in VQAv2 (Goyal et al., 2016), evaluation on VQAv2 and AdVQA (Sheng et al., 2021)) and Visual Reasoning (fine-tuning and evaluation on NLVR2 (Suhr et al., 2019)). The publicly available VQAv2 test splits do not include labels, which are required to evaluate calibration and selective prediction (*cf.* Sec. 5.2). Therefore, we follow previous work (Whitehead et al., 2022; Dancette et al., 2023) and divide the validation set of VQAv2 into dev/val/test. All results are averaged over three training runs with different seeds. Error bars show the standard error.

Hyperparameters. We use the optimal hyperparameters reported in (Kim et al., 2021; Wang et al., 2023) for AdamW. For IVON, most defaults (Shen et al. (2024)) can be used, but the learning rate and Hessian initialization need to be adjusted. However, we find that due to a strong correlation between the two, the dimensionality of the search space is effectively one. A full account is provided in Supplement Sec. A.

Sample number. Per default, Variational VQA uses $N = 64$ MC samples, as we did not find significant improvements beyond this number. For early stopping, we use eight MC samples to save compute.

[1]In selective prediction, only relative confidences matter, so there is no negative impact.
[2]As of 10/2025, see the VQAv2 Challenge on EvalAI

Temperature and Vector Scaling. Previous work (Whitehead et al., 2022) has shown that calibrating models with widespread methods like Temperature Scaling (Guo et al., 2017) and Vector Scaling (Platt et al., 1999) has only a small effect on their selective prediction performance. We confirm these findings and show that the effect is consistently positive, and can be applied on top of any method (*e.g.* AdamW or VarVQA) to receive small additional gains. Full results are in Supplement Section C.

5.2 Evaluation Metrics

Accuracy. We work with the standard VQA accuracy (Antol et al., 2015), which can also take non-integer values (0.3, 0.6, 0.9), besides 0 and 1, if less than 4 out of 10 annotators agree. NLVR2 accuracy is binary.

Calibration. We evaluate calibration using the Expected Calibration Error (ECE) (Naeini et al., 2015; Guo et al., 2017), as is standard practice. The ECE is computed by dividing the model's confidences on a dataset D into m bins D_m, and then summing the bin-wise deviations of confidence from accuracy:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|D_m|}{|D|} \cdot \left|\mathrm{Acc}(D_m) - g(D_m)\right|. \tag{9}$$

Coverage at Risk. For the selective prediction metrics, we follow prior work (Geifman and El-Yaniv, 2017; Whitehead et al., 2022; Dancette et al., 2023). The standard selective prediction metric is *Coverage at Risk* ($C@R$)[3], which measures the percentage of questions the model is able to answer (*i.e.* it does not abstain), while keeping the error tolerance r below a given risk level R:

$$C(\gamma) = \frac{1}{|D|} \sum_{x \in D} \mathbb{1}(g(x) \geq \gamma), \tag{10}$$

$$r(\gamma) = \frac{\frac{1}{|D|} \sum_{x \in D} (1 - \mathrm{Acc}(f(x))) \cdot \mathbb{1}(g(x) \geq \gamma)}{C(\gamma)}, \tag{11}$$

$$C@R = \max_{\gamma} C(\gamma) \quad \text{s.t.} \quad r(\gamma) \leq R. \tag{12}$$

We also compute the area under the Risk-Coverage curve (AUC) (Kamath et al., 2020). A weakness of $C@R$ is that the threshold γ is determined using the test set. This is necessary as otherwise, a comparison of results would be challenging: For a given risk R, one would have to judge both *threshold generalization* (*i.e.* whether the test risk matches the bound R), and the achieved test coverage.

Effective Reliability. Whitehead et al. (2022) suggested *Effective Reliability* Φ_c that avoids test set threshold selection. It differs from accuracy by a negative cost c assigned to wrong answers:

$$\phi_c(x) = \begin{cases} \mathrm{Acc}(x) & \text{if } g(x) \geq \gamma \text{ and } \mathrm{Acc}(x) > 0, \\ -c & \text{if } g(x) \geq \gamma \text{ and } \mathrm{Acc}(x) = 0, \\ 0 & \text{if } g(x) < \gamma. \end{cases} \tag{13}$$

The total effective reliability is $\Phi_c = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \phi_c(x)$, and the abstention threshold γ is determined by optimizing Φ_c on validation data. We report accuracy (Acc), $C@R$ and Φ_c per cent, while keeping the ECE in $[0, 1]$, which is consistent with Whitehead et al. (2022).

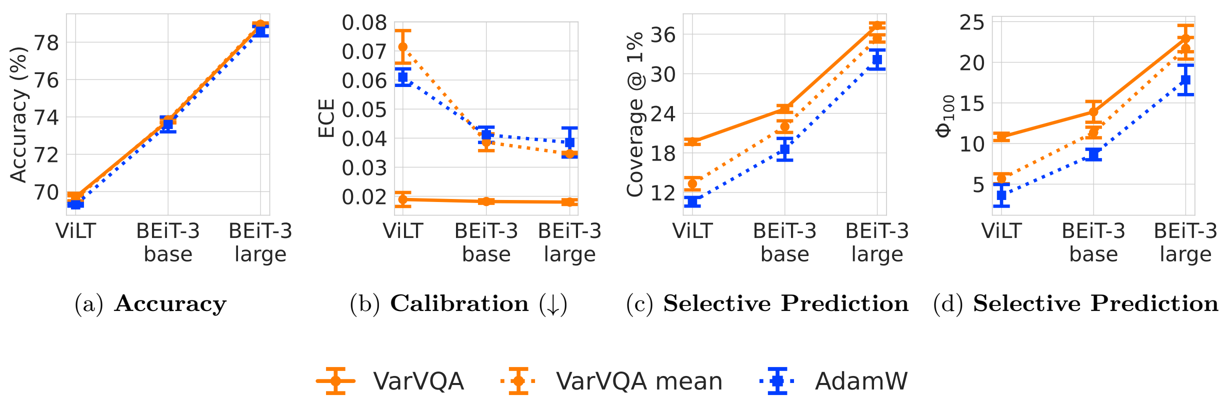
[3] A larger $C@R$ is better, as a model that abstains on (almost) all inputs is not useful.

High-Stakes metrics. Both selective prediction metrics ($C@R$ and Φ_c) feature a parameter that controls the severity of mistakes. Our findings match previous work (*cf.* Tabs. 1,2 in (Whitehead et al., 2022)): Models disproportionately struggle with settings in which errors are very costly (low-R, high-c)[4]. We collectively refer to these metrics as *high-stakes*. For practical applications, it is arguably more important that models perform well in high-stakes metrics than in low-stakes metrics, since large amounts of errors (even as low as 5%) are not acceptable in many real-world scenarios. Moreover, for ID experiments we observe saturation[5] in low-stakes metrics and thus focus our reported results on high-stakes.

It should be noted that, if stakes are set too high (*i.e.* cost c too high or risk R too small), results can become noisy, as the impact of individual overconfident samples rises. This issue increases with smaller and less well-curated datasets (label noise can have an impact). In our experiments, we observe that the results were stable only up to $c \approx 100$ and down to $R \approx \frac{1}{2}\%$, which is why we stop reporting there.

5.3 In-Distribution Experiments

We show ID results after fine-tuning on VQAv2 in Table 1 and on NLVR2 (Visual Reasoning) in Table 2. Figure 3 visualizes the VQAv2 results. As can be seen, Variational VQA matches the accuracy achieved with the conventional AdamW optimizer (Fig. 3a), indicating that Variational VQA is effective for multi-modal learning. Additionally, 'VarVQA mean' (*cf.* Sec. 4.1), which does not even use the learned posterior at inference, is frequently more reliable than AdamW (lower ECE, higher $C@R$, Φ_c), while needing the same inference compute. Finally, the VarVQA sampling strategy is the most reliable method, consistently outperforming MC Dropout, which uses the same amount of samples at inference, in terms of selective prediction, while achieving a low ECE of $\lesssim 0.03$ throughout and < 0.02 on VQAv2 with all three tested models. Regarding selective prediction, the improvements are largest for the high-stakes metrics. When only one mistake per 200 samples is allowed ($C@\frac{1}{2}\%$), VarVQA on different VLMs improves $7\% - 9\%$ on VQAv2 and $9\% - 14\%$ on NLVR2 vs. AdamW in absolute numbers.



(a) **Accuracy** (b) **Calibration** (↓) (c) **Selective Prediction** (d) **Selective Prediction**

Figure 3: Results on Accuracy, calibration and selective prediction on VQAv2 after fine-tuning.

5.4 How many MC Samples are needed?

We perform an ablation on the number of MC samples, which is directly proportional to the required inference time. Moreover, we also compare Variational VQA to MC Dropout (Gal and Ghahramani, 2016), see Fig. 4. We find that while MC Dropout often improves over the AdamW baseline, it cannot match Variational VQA in the high-stakes reliability metrics of selective prediction. For example, with BEiT-3 large, to beat Dropout@64, 2 samples are enough on VQAv2 (Fig. 4a, left) and 4 samples suffice on NLVR2 (Fig. 4b, left). Generally, Variational VQA is more sample-efficient than MC Dropout and saturates at higher reliability scores. Extended results are in Supplement Section D.

[4]The achieved $C@R$ and Φ_c in these settings are much further below the theoretical optimum than for high R/low c.
[5]For example, BEiT-3 large on VQAv2 achieves $C@10\% > 81\%$ and $C@20\% > 98\%$.

Table 1: Reliability evaluation on VQAv2 for fine-tuned models. The variable N denotes the number of forward passes. Best results per model are **bold**.

Model	Method	N	Acc.	Calibration	Selective Prediction *high-stakes*				Sel. Prediction *low-stakes*	
				ECE (↓)	$C@\frac{1}{2}\%$	$C@1\%$	Φ_{50}	Φ_{100}	$C@5\%$	Φ_{10}
ViLT	AdamW	1	69.30	0.061	5.03	10.58	8.41	2.89	36.24	24.05
	VarVQA mean	1	69.63	0.071	6.77	13.32	9.74	5.45	37.93	25.08
	AdamW Dropout	64	69.66	**0.019**	10.44	16.63	8.99	8.44	38.49	26.18
	VarVQA	64	69.71	**0.019**	**13.81**	**19.68**	**12.93**	**10.88**	**39.53**	**27.15**
BEiT-3 base	AdamW	1	73.60	0.041	10.35	18.55	15.59	8.65	47.93	33.40
	VarVQA mean	1	73.84	0.039	14.08	21.98	16.72	11.36	49.57	34.80
	AdamW Dropout	64	73.46	0.019	13.07	20.11	16.61	9.44	47.49	33.36
	VarVQA	64	73.79	**0.018**	**18.10**	**24.66**	**19.26**	**13.90**	**49.76**	**35.22**
BEiT-3 large	AdamW	1	78.59	0.039	21.63	32.15	26.31	17.80	63.19	45.83
	VarVQA mean	1	78.96	0.035	25.32	35.35	28.31	21.25	**64.83**	47.43
	AdamW Dropout	64	78.41	**0.018**	25.28	34.52	27.99	20.65	63.00	46.23
	VarVQA	64	78.89	**0.018**	**28.13**	**37.05**	**29.56**	**23.21**	64.68	**48.06**

Table 2: Reliability evaluation on NLVR2 for fine-tuned models.The variable N denotes the number of forward passes. Best results per model are **bold**.

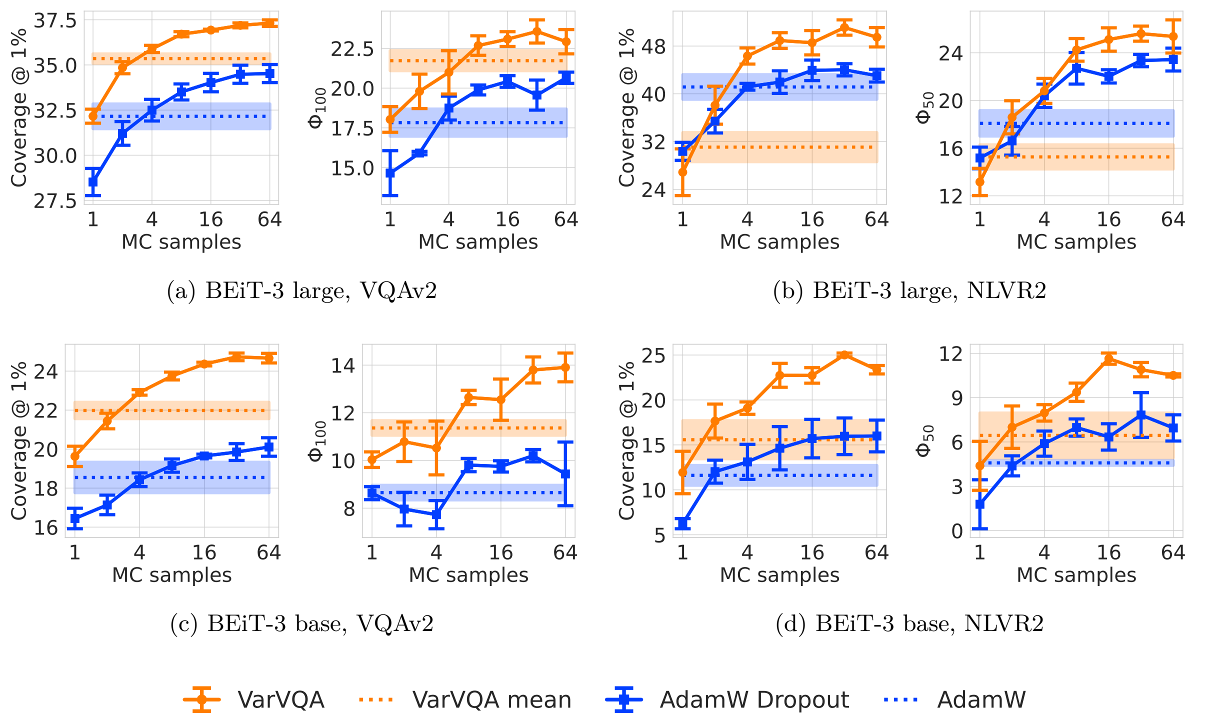
Model	Method	N	Acc.	Calibration	Selective Prediction *high-stakes*				Sel. Prediction *low-stakes*	
				ECE (↓)	$C@\frac{1}{2}\%$	$C@1\%$	Φ_{50}	Φ_{100}	$C@5\%$	Φ_{10}
BEiT-3 base	AdamW	1	83.45	0.059	6.42	11.61	4.58	2.24	54.79	26.18
	VarVQA mean	1	83.28	0.058	5.15	15.58	6.44	1.41	55.66	27.30
	AdamW Dropout	64	83.18	**0.016**	9.98	15.99	6.95	2.95	55.43	27.63
	VarVQA	64	83.11	0.031	**15.42**	**23.36**	**11.20**	**5.00**	**57.16**	**29.23**
BEiT-3 large	AdamW	1	88.34	0.041	16.53	41.14	18.08	9.45	78.53	45.64
	VarVQA mean	1	88.83	0.062	17.15	31.07	15.27	3.57	80.17	45.02
	AdamW Dropout	64	88.11	**0.017**	**33.21**	44.69	23.43	14.71	76.99	46.55
	VarVQA	64	89.26	0.029	32.89	**49.24**	**25.56**	**14.85**	**82.11**	**49.51**

5.5 Mixed ID/OOD Experiments

Following (Dancette et al., 2023), we use VQAv2 (Goyal et al., 2016) and AdVQA (Sheng et al., 2021) as ID and OOD datasets, respectively. Both datasets use COCO images (Lin et al., 2014), but AdVQA has a different multimodal distribution (more challenging questions). We use the splits from (Dancette et al., 2023), which draw testing data from P_{mix}, where

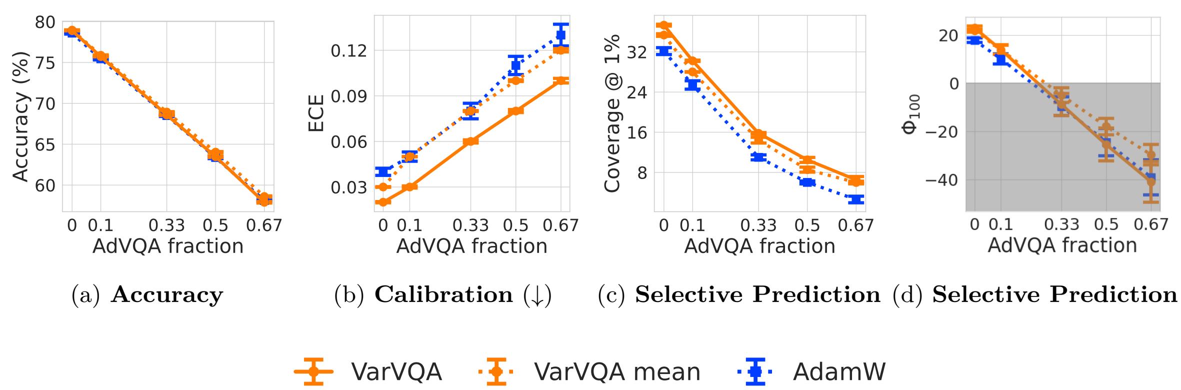
$$P_{\text{mix}} = \alpha \cdot P_{\text{OOD}} + (1 - \alpha) \cdot P_{\text{ID}}, \tag{14}$$

using P_{ID} = VQAv2 and P_{OOD} = AdVQA. Different mixtures are obtained by varying $\alpha \in [0, 1]$. Figure 5 shows the results for BEiT-3 large. Although the accuracy drops equally fast for all methods, Variational VQA remains better calibrated (Fig. 5b). The decline in $C@1\%$ is equal in absolute numbers (Fig. 5c), but this implies that the relative performance of VarVQA vs. AdamW is increasing at higher OOD fractions.



(a) BEiT-3 large, VQAv2
(b) BEiT-3 large, NLVR2
(c) BEiT-3 base, VQAv2
(d) BEiT-3 base, NLVR2

Figure 4: Comparison of Variational VQA to MC Dropout, an approximate variational method that uses the same inference compute, on the high-stakes selective prediction metrics.



(a) **Accuracy** (b) **Calibration** (↓) (c) **Selective Prediction** (d) **Selective Prediction**

Figure 5: Accuracy, calibration and selective prediction results for different VQAv2/AdVQA mixtures for BEiT-3 large. In **(d)**, every model in the gray area is worse than a model that abstains on every input.

Thus, there is reason to believe that Variational VQA may be fundamentally more robust to OOD data than AdamW-trained models. The results for the other models and metrics are in Supplement Section D.

5.6 Beyond Predictive Averaging

We compare the performance of our novel selector g_{BPA} (*cf.* Sec. 4.3) to the baseline g^{μ}_{MP} (*cf.* Sec. 4.2). The full results are shown in Tables 3 and 4. For the high-stakes selective prediction metrics, g_{BPA} consistently outperforms the sample averaging of g^{μ}_{MP}, achieving *e.g.* 5% higher $C@\frac{1}{2}\%$ on NLVR2 for BEiT-3 base. For the mostly saturated low-stakes selective prediction (grayed), there is no clear winner. When using MC Dropout, we did not find any systematic improvement of g_{BPA} over g^{μ}_{MP}.

Table 3: Ablation of our risk-averse selection function g_{BPA} (Eq. (8)) against g^{μ}_{MP} on VQAv2. Best results per model are **bold**.

Dataset	Model	Selector	*high-stakes*				*low-stakes*	
			$C@\frac{1}{2}\%$	$C@1\%$	Φ_{50}	Φ_{100}	$C@5\%$	Φ_{10}
VQAv2	ViLT	g^{μ}_{MP}	13.35	19.24	**13.04**	10.05	39.52	26.64
		g_{BPA}	**13.81**	**19.68**	12.93	**10.88**	**39.53**	**27.15**
	BEiT-3 base	g^{μ}_{MP}	17.15	23.87	18.64	12.23	**49.91**	35.17
		g_{BPA}	**18.10**	**24.66**	**19.26**	**13.90**	49.76	**35.22**
	BEiT-3 large	g^{μ}_{MP}	27.09	36.00	28.82	22.14	**64.82**	47.58
		g_{BPA}	**28.13**	**37.05**	**29.56**	**23.21**	64.68	**48.06**

Table 4: Ablation of our risk-averse selection function g_{BPA} (Eq. (8)) against g^{μ}_{MP} on NLVR2. Best results per model are **bold**.

Dataset	Model	Selector	*high-stakes*				*low-stakes*	
			$C@\frac{1}{2}\%$	$C@1\%$	Φ_{50}	Φ_{100}	$C@5\%$	Φ_{10}
NLVR2	BEiT-3 base	g^{μ}_{MP}	10.64	22.20	9.75	3.95	**57.18**	**29.28**
		g_{BPA}	**15.42**	**23.36**	**11.20**	**5.00**	57.16	29.23
	BEiT-3 large	g^{μ}_{MP}	27.61	48.16	24.26	13.59	**82.16**	**49.51**
		g_{BPA}	**32.89**	**49.24**	**25.56**	**14.85**	82.11	**49.51**

5.7 Qualitative Results

We show qualitative examples that highlight the difference in uncertainty estimates between AdamW and Variational VQA in Figures 6 and 7. Further qualitative examples for VQAv2, AdVQA and NLVR2, including failure cases, can be found in Supplement Section E. As the accuracy of the AdamW- and IVON-trained models is similar, we focus on cases where they predict the same answer, as this reflects the typical behavior. The key improvement of VarVQA lies not in better accuracy, but rather in improved uncertainty estimates. A further study that investigates the behavior on the different question categories of VQAv2 and AdVQA (*Binary*, *Number*, and *Other*), can also be found in Supplement Section E.

6 Discussion

In this work, we explore Variational VQA, *i.e.* the application of Variational Learning for multimodal tasks. Our implementation replaces the standard AdamW optimizer with the IVON method and uses multiple samples from the learned posterior at inference to achieve more reliable and well-calibrated results. Our findings demonstrate that Variational VQA has two possible applications: When inference costs should be minimal, parameter means can be used at inference to match or even slightly improve on the accuracy of AdamW and decently increase reliability. When higher inference costs are acceptable, multiple MC samples from the posterior can be used. Better reliability is demonstrated by better calibration as well as better selective prediction, both in distribution for multiple tasks, and in the challenging mixed ID/OOD setting. Moreover, we go beyond predictive averaging and introduce a novel selector function that improves selective prediction in high-stakes settings with almost no computational overhead.

Variational VQA also has some limitations, particularly involving hyperparameter tuning with IVON. While we observe correlations between the critical hyperparameters (discussed in the Supplement), which can be

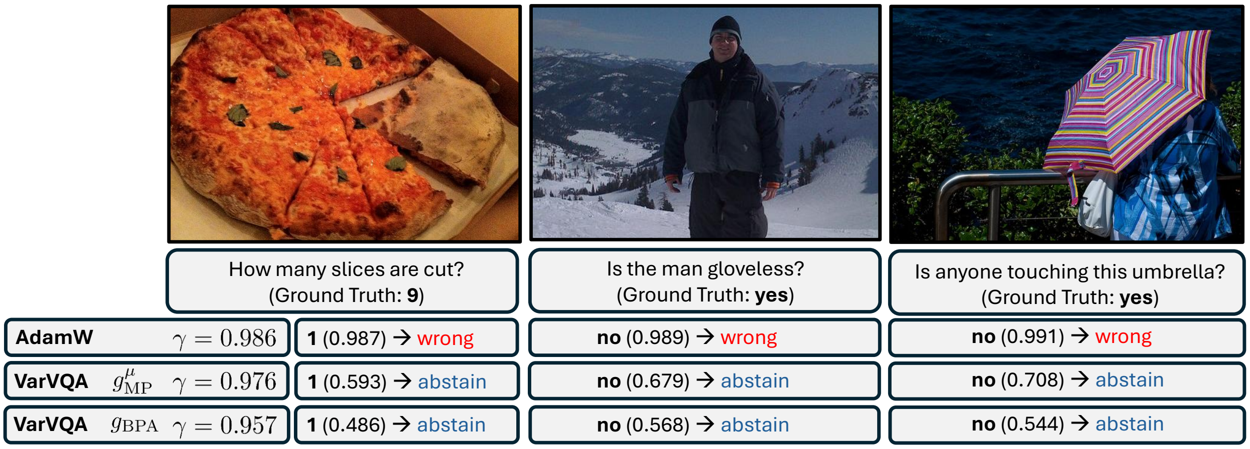


Figure 6: Qualitative examples on VQAv2 with BEiT-3 large where AdamW is wrong while VarVQA abstains. The abstention thresholds γ were determined by optimizing Φ_{100} on VQAv2 validation data. Model answers are displayed in **bold**, the corresponding answer confidences are provided in brackets.

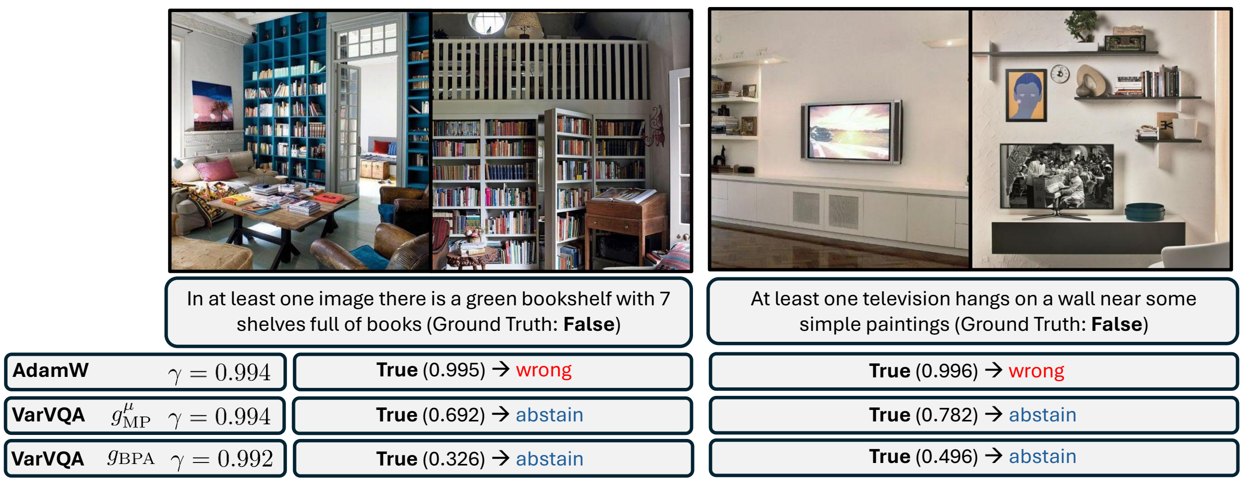


Figure 7: Qualitative examples on NLVR2 with BEiT-3 large where AdamW is wrong while VarVQA abstains. The abstention thresholds γ were determined by optimizing Φ_{100} on NLVR2 validation data. Model answers are displayed in **bold**, the corresponding answer confidences are provided in brackets.

exploited to reduce the search space, tuning still remains more involved than with AdamW. Additionally, while VarVQA makes large gains in high-stakes selective prediction vs. AdamW, overconfidence still remains an issue, and Coverages remain well below the theoretical optimum ($\approx$ *Acc.* for low risks). Thus, more work is needed to make models truly 'know what they don't know'.

An exciting avenue for future work is to avoid the computational burden of sampling for VarVQA by variance propagation in one forward pass. Recently, Li et al. (2025) proposed a new method in this domain that has shown promising results for unimodal tasks with IVON. Such 'streamlining' is only possible if learned parameter variances are available, which is not the case for *e.g.* MC Dropout. While Variational VQA intrinsically improves reliability, the incorporation of previous methods through *e.g.* training a (variational) selector on top of the (variational) model, could also further enhance reliability.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society, 2018.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, pages 2425–2433. IEEE Computer Society, 2015.

Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. Vizwiz: nearly real-time answers to visual questions. In *UIST*, pages 333–342. ACM, 2010.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *ICML*, pages 1613–1622. JMLR.org, 2015.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Ö. Arik, Tomas Pfister, and Somesh Jha. Adaptation with self-evaluation to improve selective prediction in llms. In *EMNLP (Findings)*, pages 5190–5213. Association for Computational Linguistics, 2023.

Chi-Keung Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, 4(EC-6):247–254, 1957.

Beau Coker, Wessel P. Bruinsma, David R. Burt, Weiwei Pan, and Finale Doshi-Velez. Wide mean-field bayesian neural networks ignore the data. In *AISTATS*, pages 5276–5333. PMLR, 2022.

Bai Cong, Nico Daheim, Yuesong Shen, Rio Yokota, Mohammad Emtiyaz Khan, and Thomas Möllenhoff. Improving lora with variational learning. *arXiv preprint arXiv:2506.14280*, 2025.

Nico Daheim, Clara Meister, Thomas Möllenhoff, and Iryna Gurevych. Uncertainty-aware decoding with minimum bayes risk. *CoRR*, abs/2503.05318, 2025.

Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. In *CVPR*, pages 24049–24059. IEEE, 2023.

Erik A. Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux - effortless bayesian deep learning. In *NeurIPS*, pages 20089–20103, 2021.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 11: 1605–1641, 2010.

Andrew Y. K. Foong, David R. Burt, Yingzhen Li, and Richard E. Turner. On the expressiveness of approximate inference in bayesian neural networks. In *NeurIPS*, 2020.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059. JMLR.org, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NIPS*, pages 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *ICML*, pages 2151–2159. PMLR, 2019.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *CoRR*, abs/1612.00837, 2016.

Alex Graves. Practical variational inference for neural networks. In *NIPS*, pages 2348–2356, 2011.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *UAI*, pages 876–885. AUAI Press, 2018.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. 2025.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *ACL*, pages 5684–5696. Association for Computational Linguistics, 2020.

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. In *ICML*, pages 2616–2625. PMLR, 2018.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594. PMLR, 2021.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, pages 6402–6413, 2017.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023.

Rui Li, Marcus Klasson, Arno Solin, and Martin Trapp. Streamlining prediction in bayesian deep learning. In *ICLR*. OpenReview.net, 2025.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, pages 740–755. Springer, 2014.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR (Poster)*. OpenReview.net, 2019.

David J. C. MacKay. Information-based objective functions for active data selection. *Neural Comput.*, 4(4):590–604, 1992.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022.

Erum Mushtaq, Zalan Fabian, Yavuz Faruk Bakman, Anil Ramakrishna, Mahdi Soltanolkotabi, and Salman Avestimehr. Harmony: Hidden activation representations and model output-aware uncertainty estimation for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1663–1668, 2025.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, pages 2901–2907. AAAI Press, 2015.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz Khan, Anirudh Jain, Runa Eschenhagen, Richard E. Turner, and Rio Yokota. Practical deep learning with bayesian principles. In *NeurIPS*, pages 4289–4301, 2019.

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

John W Pratt. Risk aversion in the small and in the large. In *Uncertainty in economics*, pages 59–79. Elsevier, 1978.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Yuesong Shen, Nico Daheim, Bai Cong, Peter Nickl, Gian Maria Marconi, Clement Bazan, Rio Yokota, Iryna Gurevych, Daniel Cremers, Mohammad Emtiyaz Khan, and Thomas Möllenhoff. Variational learning is effective for large deep networks. In *ICML*. OpenReview.net, 2024.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. In *NeurIPS*, pages 20346–20359, 2021.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D. Sculley, Joshua V. Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, pages 13969–13980, 2019.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Raghavi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *ACL (Findings)*, pages 12935–12948. Association for Computational Linguistics, 2024.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL (1)*, pages 6418–6428. Association for Computational Linguistics, 2019.

Brian Trippe and Richard Turner. Overpruning in variational bayesian neural networks. *arXiv preprint arXiv:1801.06230*, 2018.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In *CVPR*, pages 19175–19186. IEEE, 2023.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *ECCV (36)*, pages 148–166. Springer, 2022.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. *arXiv preprint arXiv:2505.20236*, 2025.

Liu Ziyin, Zhikang Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. In *NeurIPS*, pages 10622–10632, 2019.