We thank all reviewers for their valuable feedback. Below please find our response to each individual review.

(R1) *Significance of improvements*: For VQA, with provided error bars, the improvements are statistically significant. Especially in the more challenging noisy scenario, the improvements are over 1 point which is over 10 times the standard deviation. For NMT, the error bar for BAM-WC is 0.02 (we will add it to the revision), so the improvement is also statistically significant. Meanwhile, in Table S1, we compare the run time and number of parameters of BAM and variational attention [26], where we show that BAM achieves better results while being more efficient in time and memory.

Table S1: Step time (sec) and number of parameters of variational attention [26] and BAM on NMT.

	S/STEP	PARAMS
VA-ENUM	0.12	64M
VA-SAMPLE	0.15	64M
BAM-WC	**0.10**	**42M**

Benefits of modeling attention weights as continuous latent variables: (a) Modeling attention weights as latent variables enhances the model's ability to capture complicated dependencies and calibrate uncertainty, and prevents overfitting due to the added randomness. To evaluate the uncertainty quantitatively, we provide the PAvPU results for VQA in Table 1 (main paper) and for graph in Table S3. To evaluate it qualitatively, we visualize the predictions and uncertainties of three VQA examples in Figure S1. (b) Compared to previous work using discrete latent variables, using continuous ones is much easier to optimize. Also, BAM is faster and demands less memory and hence more suitable for low resource scenarios. On hard attention baseline, the choice of REINFORCE gradient estimator is based on previous work [9, 26]. In [26], Gumbel-softmax, which provides biased gradients, was found to underperform REINFORCE gradient estimator for NMT, which is why we have not included it in the experiment.

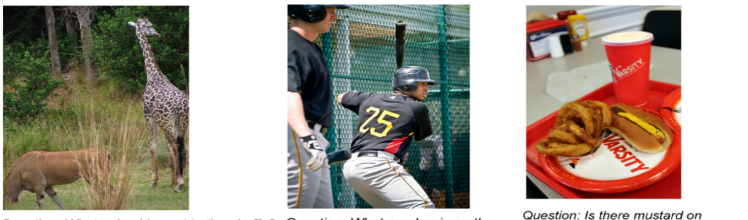

Figure S1: VQA visualization: we present three image-question pairs along with human annotations. We show the predictions and uncertainty estimates of different methods. We evaluate methods based on their answers and p-values and highlight the better answer in bold (most preferred to least preferred: correct certain > correct uncertain > incorrect uncertain > incorrect certain).

Ablation study on prior distributions: It is true that the inference network does not get the entire targets during training, which is the reason that it can be used at the test time to help predict the outputs. We introduce a contextual prior distribution to impose further regularization on the attention distributions. We agree if setting the prior and variational posterior the same, the KL in ELBO vanishes and regularization disappears. We have added an ablation study accordingly, as shown in Table S2, which suggests the importance of appropriate KL regularization. We will add them into the paper in revision.

Table S2: Accuracy for graphs.

ATTENTION	CORA	CITESEER	PUBMED
GAT	83.00	72.50	77.26
BAM (REMOVE KL)	83.39	72.91	78.50
BAM-WC	**83.81**	**73.52**	**78.82**

(R2) (1) We clarify that the proposed BAM framework works for any reparameterization distribution defined on the non-negative real line, and we have chosen Weibull and Lognormal from this family as representative examples. (2) Compared with existing method, BAM is different in not only variational distribution and gradient estimation, but also prior distribution. (3) Table S1 shows that BAM is more efficient in both time and memory than variation attention [26].

(R3) (1) Modeling attention weights is a quite standard approach [9,26,28] as they have intuitive meanings. Also in BAM, the attention weights are data dependent local variables. This approach is more computationally efficient compared to the convention in Bayesian neural network where neural network parameters, such as $\boldsymbol{\theta}$, are modeled as globally shared random variables (i.e., not data dependent), as the latter approach needs multiple sampled sets of NN weights to provide uncertainty estimation. In BAM, we only need a single set of global parameters (NN weights) and rely on the stochasticity on W to provide uncertainty. (2) We did not choose the gamma distribution as it is not reparameterizable and hence pathwise gradients that are unbiased and have low variance are not available. (3) We did not include comparison with [26] in VQA as we used a better baseline model than theirs that had already provided better performance. As their method requires a case by case design, it is unclear to us how to adapt their method to our model. Further, the code of [26] was only available for NMT so we only include the comparison for NMT. (4) In Table S3, we have included the uncertainty estimation result in terms of PAvPU for graph node classification as well and observed consistent improvements. For other tasks like image captioning, it is unclear to us how to evaluate uncertainty. (5) In the paper, we eliminated some error bar trying to prevent the table from being too crowded. We will add them in revision. *Other comments:* In our revision, we will add more detailed explanation for Figure 1, include the definition of MLE and BLEU metrics, incorporate the missing reference for evaluation metrics, and update the notation of Equation 1.

Table S3: PAvPU for graphs.

ATTENTION	CORA	CITESEER	PUBMED
GAT	82.30	72.80	77.20
BAM-WC	**83.50**	**73.90**	**78.10**