We thank all four reviewers for their candid feedback and sharp comments.

Correctness and Clarity (R1, R3) We regret our method description has confused R1 and R3, but we believe it is still mendable. Instead of directly inferring $\mathbf{z}$ from each individual query sample $\mathbf{x}$, we introduce the variational posterior $q(\mathbf{z}|S)$ in Eq. (2) and (3) to make it dependent on support set S and suitable for the few-shot setting. S infers $\mathbf{z}$ and predicts class $\mathbf{y}$ of the sample $\mathbf{x}$. $p(\mathbf{z}|\mathbf{x})$ is the conditional prior, serving only as a regularizer through the KL-term in the ELBO. R3 wonders why we do not condition the entire model on S and the variational posterior on $(\mathbf{x}_i, \mathbf{y}_i)$. Conditioning $q(\mathbf{z}|S)$ only on S fits the prototypical few-shot recognition; likelihood only depends on $\mathbf{x}$ and $\mathbf{z}$, no longer on S. $q(\mathbf{z}|S)$ is not conditioned on $(\mathbf{x}_i, \mathbf{y}_i)$ because the label $\mathbf{y}_i$ of $\mathbf{x}_i$ is unknown. To R1: latent memory $\mathbf{m}$ in Eq. (4) is indeed only for the approximate posterior, not for generative process; only support (not query) instances refer to memory (see also Fig. 1). Both $q(\mathbf{z}|\mathbf{m}, S)$; and $p(\mathbf{m}|S)$ are explicit with analytic PDFs, so we are not based on SIVI, the same for Eq. (5). The first summation in Eq. (6) is derived from Eq. (5) as follows: $\tilde{q}(\mathbf{z}|M, S) = \int q(\mathbf{z}|S, \mathbf{m}, M)q(\mathbf{m}|M, S)d\mathbf{m} = \int q(\mathbf{z}|S, \mathbf{m})q(\mathbf{m}|M, S)d\mathbf{m}$. In response to R1 and R3, all conditional distributions are implemented as uni-modal diagonal Gaussian: for $q(\mathbf{z}|S)$ we take the representation of the 1-shot sample or the average of 5-shot samples in S as input and return μ and σ of $\mathbf{z}$. Finally, the second KL-term in Eq. (13) is between the variational posterior and the conditional prior, derived from hierarchical variational inference, ensuring that the inferred latent memory is indeed relevant to the current few-shot recognition task. We will expand the method description accordingly. Thank you.

Related Works (R1, R2, R3, R4) We apologize to R2 for missing Wu et al. NeurIPS 2018. To R2 and R3, the major difference with Wu et al. (and Bornschein et al. [8] and Wu et al. [72]) is that we treat retrieved memory content – not the addressing vector – as the stochastic variable. This enables retrieved content to be better adapted to the current few-shot task. We meant that KM [72] *avoids* collapsing memory reading and writing into single memory slots, we will rephrase L50-51. We infer distributions of prototypes directly from the support sets (leveraging semantic memory), not by reconstruction as Kim et al. CVPR 2018. We will also include the other suggested references and elaborate on our relation to SIVI, Neural Process, Bayesian meta-learning and Gaussian prototypical network for few-shot learning and variational prototyping-encoder for one-shot learning.

Comparisons and Ablations (R1, R2, R3, R4) By request of R2 we add a comparison with MANN [55] (an instance of NTM for meta learning) in Table I. Indeed KM [72] and DKM (Wu et al. NeurIPS 2018) could also be explored for few-shot learning; we will implement and add them to our comparison as well. In response to R2's question on alternatives for the softmax approximation, we have implemented Gumbel-softmax. Results on *mini*ImageNet are slightly worse on 1-shot (-0.46%) and slightly better on 5-shot (+0.36%). Note Gumbel-softmax needs a predefined memory size, where our proposal can deal with dynamic memories with growing size. EMA controls how much old information will be forgotten during memory update. We choose the decay by cross validation; when varying decay from 0.8 to 0.9999 performance drops slightly (1-shot: 1.2%, 5-shot: 0.5%). We will add it to the appendix. We apply memory to conditioned $\mathbf{z}$, as already detailed in the Appendix. While $\mathbf{z}$ does not need to be stochastic, Table 1 reveals it is much better than a deterministic one. R4 asks for 20-way experiments. Besides the 20-way results on Omniglot in Table A1 of the Appendix, we report 20-way results and comparison with state-of-the-art (Jamal et al. CVPR 2019 and Chen et al. ICLR 2018) on *mini*ImageNet in Table II. We did not find prior 20-way results on *tiered*ImageNet. The requested ablation by R3 on graph attention is provided in Table A3 of the Appendix. Note our graph attention is for updating the memory, not for reading as in VMA. To R3, the improvement over ProtoNet is mainly due to the uncertainty modeling by our variational prototype net, as the major difference is modeling prototypes as probabilistic distributions rather than deterministic vectors. The probabilistic prototypes innately model uncertainty by producing distributions over predictions, which offers better confidence calibration than deterministic ones. Table 1 shows the performance benefit due to the ability of modeling uncertainty, while Fig. 2 provides the intuitive illustration. We will add more analysis on modeling uncertainty. To R1, R4, the maximum memory size is the total number of classes. We report up to the 64 classes in *mini*ImageNet. Performance could increase with more categories. Our memory is compact and has no computational burden problem. To R4, we choose sampling rates by cross validation. Indeed, larger rates can increase training time while it would make training more stable. We will add this ablation in the Appendix. Thanks for the clever suggestions to try our model on zero-shot and continual learning tasks, which we will definitely explore in future work. In response to R1, transductive BN is used for fair comparison with previous works. We will release a docker container with all source code.

Table I: Comparison with MANN (R2)

	***mini*ImageNet**		***Omniglot***	
	1-shot	5-shot	1-shot	5-shot
MANN	41.38	61.73	93.5	97.6
Ours	**54.73**	**68.01**	**99.8**	**99.9**

Table II: 20-way results (R4)

	***mini*ImageNet**		***tiered*ImageNet**	
	1-shot	5-shot	1-shot	5-shot
SOTA	19.73	38.03	n/a	n/a
Ours	**22.07**	**39.98**	**24.76**	**41.84**

We thank all reviewers for their feedback and the opportunity for reconsideration.