Author Response

We thank the reviewers for their valuable feedback. We will address the comments and the concerns as follows.

Multi-MAML vs. MMAML (R1). (1) As a whole, Multi-MAML uses all the training tasks from different modes but each of its M (the number of modes) MAML models gets trained with only the tasks sampled from the corresponding mode. MMAML does not use more data. (2) Multi-MAML assumes tasks from different modes are unrelated, whereas MMAML does not have this assumption. Therefore, we conjecture that utilizing the similarity among tasks from different modes contributes to the superior performance of MMAML. (3) As R1 suggests, there are ways to improve Multi-MAML (*e.g.* share an encoder and select dedicated classifiers for different task modes). However, Multi-MAML is not practical as it requires ground truth mode labels. Therefore, it mainly serves the purpose of verifying whether the performance of MAML degrades in multimodal task distributions. We will clarify all these points in the revised paper.

Formal Definitions (R1). Thanks for the suggestion. We agree that including formal definitions of *task*, *task distribution*, and *multimodal task distribution* would make the paper clearer and will add them in the revised paper.

Regression Training and Testing Tasks (R1). The training and testing tasks of regression experiments are sampled from the same family of functions, which is a standard setup in the existing few-shot regression literature.

Explanation on the Figure 3(a) (R1). The regression clusters overlap (*i.e.* does not reveal a clear clustering) because the observed data are noisy (with Gaussian noise) and it can be difficult to infer task modes. For example, Figure 2 of the main paper shows that the observed data of the quadratic function looks similar to a linear function. More examples of functions with ambiguity over the mode can be found in the supplementary material.

Classification Datasets (R2). (1) *CIFAR*: FC100 dataset in the paper refers to a few-shot learning version of CIFAR100 dataset (introduced in TADAM). We will clarify it in the revision. (2) *More datasets*: as suggested by R2, we replaced the two MNIST datasets with two popular datasets (CUB-200-2011 [1] and Aircraft [2]) for 5-mode experiments, similar to [3]. The results in Table R1 show that MMAML outperforms the baselines, which is consistent with finding of our main paper. We will add this to the revised paper and provide additional analysis.

Baselines (R3). We originally aimed to compare our method to the baselines that are applicable to all regression, classification, and RL. Prototypical networks, Proto-MAML, and TADAM learn a metric space for comparing samples, which are not directly applicable to regression and RL. However, we agree that it would be informative to evaluate those methods on our multimodal classification setting. We will incorporate them into the revised paper.

The code for Bayesian MAML had not been made publicly available until the paper submission deadline. During the rebuttal period, we have tried to tune and run it but have not gotten meaningful results yet. We will further consult with the authors in order to get it working so that we can add it to the revised paper.

Modulation and Adaptation (R3). We agree that analyzing the effect of modulation and adaptation steps would be helpful. However, the norm of the modulation step (τ) and the norm of the adaptation step (in the parameter space: θ) are not directly comparable. As an alternative, we provide an additional analysis showing the effects of modulation and adaptation qualitatively (shown in Figure R1) and quantitatively (shown in Table R2) by testing a trained MMAML in the 5-mode regression task and measuring the MSE of each step. Note that MMAML starts from a learned prior parameters (denoted as *prior params*), and then performs modulation and adaptation steps.

Table R1: 5-mode Classification: Omniglot, Mini-ImageNet, FC100, CUB, and Aircraft.

Setup	5w1s	5w5s	20w1s
MAML	44.09%	54.41%	28.85%
Multi-MAML	45.46%	55.92%	33.78%
MMAML (ours)	**49.06%**	**60.83%**	**33.97%**

Figure R1: 5-mode Regression: Visualization with Linear & Quadratic Function.

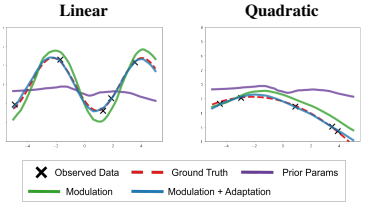

Table R2: 5-mode Regression: Performance measured in mean squared error (MSE).

MMAML	**MSE**
Prior Params	17.299
+ **Modulation**	2.166
+ **Adaptation**	0.868

References

[1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 1

[2] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint airxiv:1306.5151*. 1

[3] E. Triantafillou, T. Zhu, V. Dumoulin, P. Lamblin, K. Xu, R. Goroshin, C. Gelada, K. Swersky, P.-A. Manzagol, and H. Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *Meta-Learning Workshop at NeurIPS*, 2018. 1