We thank reviewers for detailed and helpful reviews. We particularly acknowledge that reviewers find this work effective (R1, R2, R3), well written (R1, R2), interesting (R3) and achieves good results (R1, R3). Next, we address the main concerns from reviewers.

**Presentation and extra ablation studies (R1)**. We thank R1 for pointing some expositions issues and the proposed improvements will be added to final version of the manuscript. We conduct 10-way-1-shot, 10-way-5-shot experiments on *mini*ImageNet to verify that the performance improvement of AM3 is consistent when the number of classes change. Table 1 shows the results. Due to time limits, we only compared AM3 with its backbone methods and the modality alignment method that performed the best in 5-way scenarios. Results show that AM3 also outperforms its backbone methods and the modality-alignment baseline in 10-way setting. After rebuttal, we will test all baselines on *mini*ImageNet 10-way setting to make a thorough comparison and try more variance of N-way FSL settings.

| Model | 1-shot | 5-shot |
|---|---|---|
| CADA-VAE-FSL | 37.5% | 56.3% |
| ProtoNets++ | 39.1% | 59.5% |
| AM3-ProtoNets++ | 45.7% | 61.4% |
| TADAM | 42.7% | 61.2% |
| AM3-TADAM | 47.3% | 62.1% |

Table 1: 10-way classification accuracy.

**Realistic scenario (R2).** We argue that cross-modal few-shot learning scenario (ie, having access to both train and test word embeddings) is realistic. If we understand correctly, R2's main concern is that the word embeddings of the test labels may not be accessible. We believe that it would hardly happen. The reasons are as follows.

First, GloVe (chosen as pretrained word embeddings) is trained *unsupervised* and contain embeddings for 1.9M words. It is likely that a semantic label will lie in the GloVe vocabulary. Even if it is not the case, it would still be realistic to simply crawl considerably amount of text from the web that contains the token of the test label, to (unsupervised) train a word embedding model using the same technology. Therefore, as long as we have access to the labels of the test set, getting meaningful word embeddings for them without any supervision (human labeling efforts) is relatively straight forward. Second, we can easily assume a FSL scenario in which we have access to the labels of the test set. In vanilla FSL, a support set of "few-shot" samples is provided for each unseen category. In our scenario, we assume the label of each support set is also given (eg, images of cat and the semantic label 'cat'). We found this a realistic assumption. Third, leveraging word embeddings (trained on unlabeled text corpora) of class labels (for both train and test classes) for vision tasks has long been exploited (eg, ZSL, image retrieval, image captioning, etc.)

**"Method is a simple extension of prototypical networks" (R2).** We disagree. On the contrary, AM3 is model-agnostic to any metric-based FSL methods, as described in the paper. In experiments, we test AM3 on two different metric-based FSL: ProtoNets and TADAM. We agree with R1 that "the fact that the method is agnostic with respect to which metric-based model it is extending is a positive". Moreover, we think the "extension" is not simple. Although employing extra knowledge source to help FSL may sound straightforward, the cross-modal method should be designed to fit FSL scenario – a task that is not trivial. Existing complicated cross-modality models (modality-alignment methods and proposed baselines) fail to work well in FSL, as our experiments show. The main contribution of our paper is the model that is designed to conduct cross modality specifically in FSL scenario. We empirically demonstrated it to be effective at integrating extra information from unsupervised text corpus to boost performance on the few-shot image recognition task.

**Simplicity of the model (R2).** R2 points the simplicity of the model as a weakness. We share the opinion of R1 and R3, mentioning this work is "simple yet effective" and "interesting and effective". Most of the modality alignment baselines we compared are quite complicated (eg, CADA-VAE employs 2 VAEs). However, due to their assumption that the two modalities have to be aligned (too rigid for FSL, as argued on the paper), their performances can't outperform AM3. A model as simple as AM3 outperforms complicated baselines to a large margin. In this circumstance, we disagree that the simplicity of AM3 is its weaknesses.

**Application beyond image classification (R2).** As pointed by R1 and R3, the proposed approach can potentially be used in many different cross-modal FSL settings involving visual and semantic information. Few-shot semantic segmentation, object detection, action recognition, etc, can be some of them.

**Gated formulation in Eq. 5 (R3).** We thank R3 for the suggestion on the input of the gated formulation in Eq. 5. We agree that intuitively it would make more sense if $\lambda$ is conditioned by both variables. We will empirically verify it after the rebuttal and update the paper accordingly. We will also discuss the differences wrt the papers mentioned by R3.

**"Any other work having similar ideas or implementations?" (R3).** To the best of our knowledge, AM3 is the first model in FSL setting that proposes a gated fusion of representations of the two modalities. Other models that incorporate cross-modal information in low-data regime (eg, ZSL and FSL) are based on modality-alignment methods. As argued on the submission, modality-alignment methods force the two spaces to have the same semantic structure, which is too rigid for FSL, given that we have supports from the original modality at test.