We would like to thank the reviewers for their detailed comments and feedback. All new experiment results have conducted on the MS COCO detection dataset (Sec. 4.1.1).

R1,4. Prior Work. The state of the art approach for the partially annotated multi-label classification task is [14]. There are two contributions of [14]. One of them is the nWE baseline, which our approach outperforms. The other contribution is using GNN. But it barely has any improvement. In our settings we do see a similar trend of achieving $< 0.2\%$ improvement in mAP. From a high level, instead of modeling label relations directly from the data, we use priors in terms of distance of class embeddings. Moreover, we exploit image similarities as well in this approach.

R1,2,3. Sensitivity to hyper-parameters. We selected the values of $\beta = 5$ and $\gamma = 0$ based on the validation set. The mAP is within 1% of the reported performance (on average) for $\gamma \in (0, .1]$. The drop in performance can be as large as 5% with $\gamma \in (0, 1]$. For,$\beta \in \{1, 2, 5, 10, 20, 50, 100\}$, avg. mAP on the validation set was within 2% of the best performance at $\beta = 5$. For values of $5 < k \leq 30$, the SEI performance increases by 5%, but it's improvement on the SE model is $< 0.15\%$. $k < 5$, reduces the performance of SEI and SE models and brings them closer to NE and SEL models respectively.

R1. Design choices of SE modeling. The main motivation of the paper is to use image-image and label-label relationships to capture more supervisory signal from the unsupervised un-annotated labels. We implemented this via temperature modeling. Exploring better modeling choices is a work in progress. Thank you for suggesting the entropy based modeling. Regarding the "hard" minimum operations, we also experimented with "softer" operations in this rebuttal. Instead of taking the minimum, we take the median operation in Eq. [4] and [5] among the top 5 neighbors. We see an improvement of ∼0.7% for label-label relationships using this approach.
While both label and image based relationships improve the performance, we do observe that the label based distances dominate the image based. This is because the number of labels considered for the image based distances is significantly lesser than the label based distances. While additional 72.7 labels are considered for the label based, the number of labels being considered for image based is ~ 5.5 @10% partially annotated data. We will perform an in-depth analysis of the effect of this discrepancy on our approach.
Initialize embeddings on LS baseline. This improved our performance by upto **1.5%** mAP.
Distance computation cost. It takes <1 epoch training time (∼15 min. on a single V100 GPU) and it's done once.
Paper improvements. Thank you for the feedback. We will improve the writing of the core section as well as the visualization of Fig. 5 in the camera ready draft.

R2. Feature pre-processing for distance computations. We process the features in the same way as [72], where we use the 2048-dim feature vector and do L2 normalization on them. For k-NN, we use these features to compute the neighbors. For $\psi(c)$, we take a median of these representations across all images where, c occurs. We had also experimented with mean, but found median to have better performance.
d_L distances when $P(x), N(x) = \phi$. Implementation-wise, we ignore such labels when this happens. However, when combined with d_I, the overall distance value defaults to 1 based on Eq. 5.
Validity of results on Multi-Label CIFAR dataset. As rightly pointed out, CIFAR is indeed a single label dataset and the multiple labels is created because of the hierarchy of the knowledge graph. The purpose of this dataset is to explore the effectiveness of this approach when there is a single object visually present in the image. However, we experiment with other multi-label datasets such as MS COCO detection and panoptic segmentation, and real-world partially annotated multi-label datasets such as OpenImages and LVIS.

R3. Inconsistency of results with [14]. Compared to [14], the main difference is that [14] uses an older split of MS COCO training and validation set (which was taken from an older paper), which is not considered standard in the object detection and multi-label classification literature anymore. We used the training setup of [66] for our experiments. The oracle (with 100% labels) results match that of [66]. We had ran our approach using the setting mentioned in [14], and can conclude the same trend as observed here. We will add these results in our final draft.
Single label performance. Our SE model improves the best performing FE baseline by **13.5%** in mAP.
Absence of pre-trained networks. This is a great question. In a trivial way, we can compute similarities after every "few" epochs. Initially, we can simply use the LS modeling. But we can investigate this issue further and explore meta learning or more sophisticated approaches.
Missing Reference. Thank you for the missing reference. We will add it in our final draft.

R4. Classification vs Detection tasks. Multi-label classification is a well-established task and MS COCO along with OpenImages are considered standard benchmarks for this task. Detection is another task, which along with image-level labels, also require bounding box annotations for localization.