We individually respond to the questions and concerns raised by reviewers about our work below:

Reviewer 1:

- **"The framework is natural but straight-forward"**: We are glad that R1 thinks that our framework is natural. However, inspite of being simple, such a unified view has not been applied to explain popular representation learning approaches, especially recent ones such as masked self-supervision and VAE. We believe that formulating these approaches theoretically and analysing their sample complexity bounds is an important problem in the domain of representation learning and can lead to insights. We view the simplicity of the problem formulation to be a strength rather than a weakness.
- Additionally, we present two concrete applications of the framework, with empirical support thereby showing the validity of our framework. This shows that our general framework is effective and doesn't lead to vacuous bounds.

Reviewer 2:

- **"Perhaps to have more focus on cases when this will not work"**: Our theorems and results do provide implications for cases when the auxiliary self-supervised task may not help the prediction task. For example consider Theorem 1, if $\mathcal{H}_{\mathcal{D}_X,L_r}(\epsilon_0)$ is not significantly smaller than $\mathcal{H}$, then using unlabeled data will not reduce the sample size of the labeled data much compared to only using the labeled data for prediction. Further, for the two concrete examples that we present in Section 5, the sample complexity bounds also indicate when the unlabeled data will not be very useful. Another possible reason that these representation learning approaches can fail is that the optimization does not provide a good solution. However, accounting for this is out of scope of this work. We will incorporate the suggestion and add more discussion about the related works pointed out.
- **Clarification about Equation 15**: This equation means that if we fix a pair (h, g) with $L_r(h, g, D_x) \geq \epsilon_0$, then we have $P(L_r(h, g, U) = 0) \leq \delta/2|H||G|$. Since there are at most $|H||G|$ such (h, g) pairs, by the union bound we have $P\big(\exists(h, g)$ s.t. $L_r(h, g, D_x) \geq \epsilon_0$ and $L_r(h, g, U) = 0\big) \leq \delta/2$. Then with probability at least $(1 - \delta/2)$, there exists no (h, g) such that $L_r(h, g, D_x) \geq \epsilon_0$ and $L_r(h, g, U) = 0$. This means that only those (h, g) with $L_r(h, g, D_x) \leq \epsilon_0$ will have $L_r(h, g, U) = 0$.

Reviewer 3:

- **"Interaction between the two learners"**: Our sample complexity bounds do quantify the interaction between the two learners (from the hypothesis classes $\mathcal{H}$ and $\mathcal{G}$) by introducing the notion of $\mathcal{H}_{\mathcal{D}_X,L_r}$. This captures the effect that the representation learner over $\mathcal{G} \circ \mathcal{H}$ has on the prediction learner over $\mathcal{F} \circ \mathcal{H}$.
- **"Covering numbers are not very effective in practice"**: While we acknowledge the evidence that naively applying uniform convergence bounds may not result in good generalization/sample bounds for deep learning, these existing studies may not apply to the setting we consider in our paper. To the best of our knowledge, this evidence is specific to supervised learning without the auxiliary representation learning tasks, while our setting is with the auxiliary tasks. In particular, the mentioned paper "Uniform convergence may be unable to explain generalization in deep learning" does not apply directly to VAE or any other representation learning approach studied in our work.

  Furthermore, our experimental results do correspond with our sample bounds in Section 5 which are based on uniform convergence. This is in contrast to the existing studies on supervised deep learning without auxiliary representation learning tasks. These results suggest considering a shifted view: "Uniform Convergence strikes back and can explain the generalization behavior of deep learning *with* auxiliary representation learning tasks". Why so? Without auxiliary representation learning tasks, it is generally believed that the optimization has an implicit regularization on the training, and hence uniform convergence fails to explain this. However with the auxiliary tasks, we conjecture that functional regularization restricts the learning dynamics to a smaller subset of hypotheses, on which the implicit regularization of the optimization is no longer significant, and thus the generalization can be explained by uniform convergence. This is an interesting open question and we leave it as future work.

Reviewer 4:

- **"Experiments on real world data would have helped"**: We have presented some experiments on real data in the appendix. Due to space restrictions, we only focus on experiments for the two concrete instantiations of our framework, which we believe can give fine-grained empirical evidence (e.g., how the sample bounds depend on r, etc.) for our analysis. We will move some experiments on real data to the main body in a future version of our paper.
- **Line 135**: The loss $l_c(f(h(x)), y)$ denotes the loss function for the prediction task, and has been defined on line 90.
- **Line 179**: By "standard analysis", we refer to the standard statistical learning theory argument for uniform convergence over a finite hypothesis class. We will make this explicitly clear in in a future version of our paper.
- **Related work**: Thanks for the suggestions! We will add them to the future version.