We thank all the reviewers for their insightful comments and suggestions that will all be incorporated in the final version. We start our response with Reviewer #3, since some of the comments made there are inaccurate.

Response to Reviewer #3: The major critique is that our paper rediscovers prior ideas. However, the two papers of Ben-David and others pointed out by the reviewer are essentially different in the setups and results, which is also why those papers were not cited in more related works we cite (e.g., [11,17,31]) either. Below we expand these differences:

1. The setting. They assumed a *multitask learning* setting where n tasks/datasets symmetrically contribute to the loss function, and the learner has access to *datasets for all tasks* (i.e., theorem 3.5 assumes $|S_i| \geq m$ for all i). In contrast, our framework assumes an asymmetric *supervision* setting where the learner only has access to samples of (X, O) but *no dataset for the target label* Y. Moreover, They assumed that all tasks are *binary classification* while in our framework $\mathcal{Y}$ and $\mathcal{O}$ can be *multiclass* and different. Also, they define the relatedness of tasks by deterministic *functions* $X \to X$ while in we define it as *conditional distributions* $\mathbb{P}(O|Y)$.
2. The analysis and results. In their work, since all tasks are associated with a dataset of size $\geq m$, the generalization is given for-free, as long as the generalized VC-dimension is finite. However, in our paper, learnability is nontrivial (e.g., think about a very noisy dataset). So, our paper makes additional effort to characterize the transition class and the learner's prior knowledge about it, which results in conditions [C2] and [C3].

The other critique is we make strong assumptions. In fact, our assumptions are *common* and often *weaker* than literature.

1. Realizability assumption. This is commonly assumed in the related works (see line 34) and is approximately satisfied by many practical models (such as DNNs that achieve classification error close to 0 on many large, practical problems). Due to the stochastic generating process of the indirect signal, it would be technically hard to remove this assumption in a general setting. We therefore think that it is a good starting point for developing the theory.
2. Learnability conditions. First, the conditions [C2] and [C3] actually generalize and relax many standard learning conditions in the literature (see Example 5.6 & 5.7). Moreover, the second part of our theorem 5.2 shows that the conditions are not only sufficient, but also close to be necessary. In other words, our framework provides more practical (rather than unrealistic) ways to determine learnability and/or find learnable supervision signals.

Response to Reviewer #1: Here are our responses to the concerns:

1. Use of VC-style argument. We choose to use VC-style argument to bound the Rademacher complexity because it is a *safe* way to show *learnability*. It is safe in the sense that most of the practical machine learning models, including DNNs, have a finite VC-dimension. Also, it is possible to adopt to a PAC-bayesian argument by defining induced prior and posterior via the transitions. However, to show learnability in this way, we would need to further bound the KL-divergence by constructing priors for specific models. It would be an interesting direction for future research.
2. Practical implications. We will address this issue more clearly in the final version. Briefly: To make full use of the indirect observations, our framework provides a practical and unfied way to understand the supervision power of the signal when only weaker form of the prior knowledge is given (encoded by *separation*). Our framework generalizes previous results, and hence suggests new learning scenarios. Also, in a non-learnable case, section 5.3 also guides to find complementary signals to make the problem learnable.

Response to Reviewer #2: We thank the reviewer for the useful feedback. For the weakness and additional feedback:

1. To derive a matching lower bound, we will need two things: (1) For the case when $\mathcal{T} = \{I\}$, it is true that faster rate can be achieved. This is partly because the task $X \to O$ is agnostic in default. So if using the current method, (at least) we will need to define what is a "realizable" O. (2) A new measure similar to separation that links the annotation and classification risk. We will add more discussion on this issue and/or the limitations fo current work.
2. We will add an illustrating example for this phenomenon 199-202. For linear case, the intuition is that different regions will have different patterns of the distribution of O. We will also add a visualization to make it clear.
3. For the additional feedback: **(1,5)** There are typos. Will be corrected in the complete version. **(2)** We will adopt the suggested way of definition and call it "$\mathcal{T}$-learnable". **(3)** b is the upper bound of the annotation loss $\ell_{\mathcal{O}}$, which may not be the 0-1 loss. We will add a reminder here. **(4)** We will add a definition for the Natarajan dimension. **(6)** We can restate the definition as: $\mathcal{D}_i(x)$ is the set of all the induced distributions $\mathbb{P}_T(O|X = x, Y = y_i)$. **(7)** $\dim$ is used to indicate the singularity of T, since one may concern if a singular T will lead to information loss. **(8)** Our framework can be extended to product case and the product label will contain more information and provably have better supervision power. However, to formulate this idea in general, we also need to model and characterize the *correlations* between different annotations (which could be expensive). It will be an interesting future research direction. **(9)** Your intuition is correct. We would describe it like this: [C2]: the optimal classifier of Y can induce an optimal predictor of O. [C3]: the suboptimal classifier of Y will induce a suboptimal predictor of O.