Many thanks for the precious comments. And we really appreciate the recognition to the contribution of this work!

To Reviewer#1

Q1: Some more ablations. **A1**: Table 1 shows the ablation study for the second stage on localization loss $\mathcal{L}_1$, sounding area suppress operation *Prod* and category matching objective $\mathcal{L}_c$. Our model achieves the best when with all of them. **Q2**: How long for training the first stage vs the second stage? **A2**: Both two stages require about 10k training iterations to reach the best performance.

To Reviewer#2

Q1: The information about single/multiple sources. **A1**: The information comes from the dataset itself, e.g., solo/duet partition in MUSIC and number of annotated sound source in AudioSet. However, such strict partition does not always satisfy. We found that the AudioSet videos annotated as single source usually contain multiple sounding objects, e.g., *-g4i39nadkQ* and *MeGMI06BmLs* (YouTube video ID). Even faced with such noisy and incorrect partition, our method still generalizes well and provides good object representation for the second stage learning.

Q2: Number of object categories. **A2**: The number of object categories is equal to instruments in dataset for easy evaluation. However, Table 2 shows different cluster numbers, and our model still achieves comparable results.

Q3: Using softmax over all object categories? **A3**: 1. Using softmax can simultaneously encourage the model to find sounding objects and suppress the silent ones. 2. The followed KL divergence requires the input to be a distribution.

Q4: Using the balanced instead of the eval set in Audioset for testing. **A4**: For fairness, we just follow the previous works and use the balanced set for testing. As the balanced and eval set are constructed with same rules, the testing performance should be consistent. If permitted, we will append the comparison results on the eval set.

Q5: The influence of threshold for binarization for mask. **A5**: Table 2 shows the influence of mask threshold, 0.03, 0.05 and 0.07. Our method is robust to the choice of threshold.

Q6: Generalization on universal objects. **A6**: 1. Playing instruments is a typical audiovisual scenario, and generally used for evaluation previously [3, 30]. 2. We partially examine generalization on the challenging YouTube video (AudioSet). Although it also consists of playing instruments, these wild videos are very noisy, many are of poor quality and mixed with other object sounds [3]. 3. Although lack of benchmark dataset, we agree that the more general daily-object should be explored. In future, we will focus on these to further facilitate the development of sound localization.

Q7: Localization for occluded objects. **A7**: We show localization results for occluded objects of guitar and violin in Fig. 1, and we think our method is robust to partial occlusion as long as some key parts of objects are exposed.

To Reviewer#3

Q1: Whether the labels there are only for illustrative purposes? **A1**: The label in Fig.4 is only for illustrative purpose and not used in the training stage. The details about semantic label acquirement can be found in the sup material.

Q2: Number of object categories. **A2**: We explore using different number of clusters in training in Table 2, our method is robust to the cluster number. Hence, we do not certainly require it in training, but use it for evaluation.

Q3: The novelty/contribution. **A3**: 1. Unlike the simple audiovisual cases in previous works, this work deals with a more realistic and complicated cocktail-party scenario (confirmed by R#1 and R#4), meanwhile targets to discriminatively localize sources without manual semantic label. 2. The proposed techniques, including robust object representation learning, self-supervised category matching, step-by-step learning paradigm etc, are obviously different from previous localization approaches and recognized by R#1 and R#2.

Q4: Noisy single source video. **A4**: Please see A1 to R#2.

To Reviewer#4

Q1: Quantitative ablation study in Eq. 7. **A1**: In Table 1, without the consistency objective $\mathcal{L}_c$ but only scene-level correspondence, there is no supervision to facilitate discriminative localization, causing the dramatic performance drop.

Q2: Temporal information for localization. **A2**: Yes! Temporal information can be considered as a potential cue for sound localization, and we plan to explore this in future. Many thanks for the comment on such promising direction.

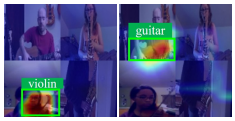

Figure 1: Occluded obj.

Table 1: Ablation study for the second stage.

Dataset			Music-Synthetic		
$\mathcal{L}_1$	Prod	$\mathcal{L}_c$	CIoU	AUC	NSA
✗	✗	✓	0.0	7.2	91.0
✓	✗	✓	2.6	7.5	88.1
✓	✓	✗	18.0	17.4	92.9
✓	✓	✓	32.3	23.5	98.5

Table 2: Ablation on threshold, cluster number.

Dataset		Music-Synthetic		
Threshold	Cluster	CIoU	AUC	NSA
0.05	11	32.3	23.5	98.5
0.03	11	31.4	24.2	96.2
0.07	11	32.1	24.0	94.5
0.05	13	33.8	24.0	96.2
0.05	20	29.5	22.2	98.9