We thank all the reviewers for your careful reading of this paper and constructive comments. We did our best to answer all of the questions and supplement the experiments as much as possible in this limited rebuttal period.

For Reviewer 1: **(3.1)** We conducted speech translation experiments on the Fisher and CALLHOME Spanish-English corpus with the proposed conditional model. Although we could not finish the entire training and tuning hyper parameters due to the limited time, we obtained promising results on a validation set at the 12th training epoch (BLEU: 67.2), compared with the Transformer model (BLEU: 63.1), with the same configuration. We will add the experiments to Section 5 to provide more evidence about the generalization of our work. **(3.2)** As you suggested, we trained a conditional ASR model with a similar configuration for Method 7 in Table 3, except that the labels and conditions are randomly ordered. The WER of random ordering is 28.4%, worse than the greedy method 24.7%. **(3.3)** We fully agree with this point and we will apply our method to more realistic scenarios, e.g. CHiME6 and CALLHOME, as the most important future work. We are confident that our method can be used in such scenarios, since the related study of our conditional model in speaker diarization actually [10] got very positive results in the CALLHOME task. **(6)** The two papers you mentioned use a different dataset (based on **WSJ1**, which is *five-times larger* than **WSJ0**) and we cannot directly compare them. Instead, we performed the experiment with the exact same WSJ0-2mix data by using our reproduced version of [6] (Method 5 in Table 3), which is similar to the first work you mentioned. We will clarify this point in the manuscript. **(8)** As you suggested, we trained our conditional model with zero initial LSTM hidden vector for each sequence in the decoder with the same configuration as Method 7 in Table 3. The WER on WSJ0-2mix was degraded by 8.6%, which proves that the whole history information does help. This result will be added to the paper.

For Reviewer 2: **(3)** We really appreciate your comments about the relationship with several related studies. We fully agree with your point and we did not intend to claim that the conditional methodology itself is novel. Instead, we wanted to claim that the novelty of this paper is to apply such a concept to analyze "Mixture Signals" (as shown in the title and the other part of the paper) especially when the number of sources is not known. This often happens in audio/speech signals, especially for multispeaker scenarios, as you pointed out. We tried to clarify this point during the submission stage, but we will further clarify this point in our abstract and introduction. In addition, we re-trained our model in the spatialized wsj0-2mix dataset, compared with the same configuration and network from an open-source implementation with 8-ch MVDR beamformer in the time-frequency domain. Our results on the 2-mix reverberant condition get similar results with the baseline (9.57db vs 9.78 db SDR) without much tuning, which proves the adaptability of our method even under reverberant conditions. We will continue these experiments and add the results into our appendix part. **(8)** Thanks a lot. We will polish the article with your suggestion. (Sec 3.2) For stop criterion, we use average energy to determine pure silence in separation task (see line 241) and average posterior of <blank> label in ASR task. (Tab. 3) We agree with your point and applied the 1,2,3Mix utterances into the training corpora for baseline models. As you pointed out, this additional experiment shows the similar performance improvement (19.2% for 2mix and 39.0% for 3mix), but the proposed model is still better (14.9% for 2mix and 37.9% for 3mix), which will be added to Table 3.

For Reviewer 3: **(3.1)** Because of the chain design of our model, we could make better use of different datasets (e.g. with a different number of speakers). Experimentally, our model does improve from the mixing of multiple data sets, but we also achieve better results for a single data set than the baseline approach (see in Table 1&2). This proves the effectiveness to use the conditional chain to fit the relationship between multiple outputs even for a single dataset. **(3.2)** Yes, one of our future work is to introduce the attention mechanism into the fusion and conditional chain part, which could flexibly capture the implicit relationship between input and output from variable domains, as we mentioned it in Section 6. Actually, we already applied it to speech translation tasks, and it seems to be working (See Reviewer 1(3.1)). **(8.1)** and **(8.2)** Currently, our method is only applied to the near-field condition, but some extended experiments (e.g. spatialized WSJ0-mix, see in reply for Reviews 2 (3)) have been done and showing promising results under the reverberant, far-field, and various mic/source location conditions. Our ongoing work is to apply it to more realistic CHIME-5/6 datasets, as you suggested. **(8.3)** Yes, a word-LM is used. This is described in Section 5.3 first paragraph, and we will rephrase this paragraph to clarify it. **(8.4)** In all our experiments, the model seems to be quite stable. We did experiments in several different environments, and the problem of repeatability did not arise.

For Reviewer 4: **(3.1)** We appreciate your comments about the lack of clarification about the different training sets. We agree that Table 1 is misleading and we'll add more detailed information about the training data. Also, we are working on the fair comparisons with DPRNN (see the discussion below). **(3.2)** We agree with your point. Among Voice Separation and DPRNN, we actually picked up DPRNN in addition to TasNet, and performed experiments based on the conditional-DPRNN model during the submission stage. The reason we did not list the DPRNN result is that our implemented DPRNN could not reproduce comparable results on WSJ0-2mix (18.0 vs 18.8 dB SI-SNRi in [24]). However, the results with our implementation also shows a similar tendency (e.g., the conditional-DPRNN is 0.2 dB better than the DPRNN in WSJ0-2mix and the WSJ0-2&3 training gets further improvements of 0.3 dB. We can add our DPRNN result to the paper once we reproduce the comparable numbers in [24]. **(3.3)** We will add the explanation of SI-SNR accordingly. **(3.4)** Thanks for pointing out this. We conducted experiment on the baseline model with 1, 2, 3Mix utterances. Please check our reply for Review 2 (8, Tab.3).