We thank all the reviewers for their constructive comments. Below, we address concerns raised by reviewers.

In-the-wild experiments (R1, R3). Because of our insufficient domain knowledge on German signs languages, we opt to collect 76 short videos in German sign language from the internet as in [10]. In this way, we ensure our model trained on RPWT fully covers the vocabularies, thus facilitating the validation of translations. Results show that the proposed TSPNet correctly translates 32 phrases despite the domain gap. Although there are a large number of sign videos, annotating them with glosses requires expert knowledge while being time-consuming and laborious. Without such gloss annotations, our method is still able to construct a sign language translation model. Moreover, the data for training our model are easy to collect similar to [10]. These reasons make our method more favorable.

Discussions on Cihan et al., 2020 (R1, R3). Cihan et al. employ a visual encoder finetuned on the RWTH-PHOENIX dataset (GSL) *with gloss annotations*. Training/finetuning on gloss annotations will facilitate a model to effectively infer sign boundaries. In other words, Cihan et al.'s method requires gloss annotations of a particular dataset to achieve its best performance. Thus, comparing our results with theirs is unfair since we do not use gloss annotations.

Current limitations (R1). (i) We notice that low-frequency words, such as city names, are very challenging to translate, we plan to provide our models with an external knowledge database and allow it to search similar gestures for translation. (ii) Facial expressions often reflect the extent of an event, e.g. shower *vs.* rain storm, which are currently not well interpreted by our models. Our future work will try to incorporate facial landmarks to improve translation accuracy. We will include the above discussions in the revision.

Alternative visual backbone and language decoder (R2, R3). (i) TSPNet aims at proposing a generic learning framework to translate continuous signing videos to natural language sentences directly. We adopt I3D as visual feature backbone considering its recent success in sign language interpretation [8, 9, 10]. As SlowFast network demonstrates its superior action modeling capacity to I3D, we believe employing a stronger visual backbone, like SlowFast network, would lead to better translation results. Due to the tight timeframe of rebuttal, we cannot report results using the SlowFast network. We will add the results and discuss the impact of alternative visual encoders in the revised version. (ii) As suggested by R3, after replacing a Transformer decoder with an RNN in [2], we did not observe performance changes in BLEU-4 (+0.09) or ROUGE-L (-0.03). This implies that the performance gains mainly come from the proposed hierarchical encoding scheme rather the decoder.

Clarification on backbones (i) (**R2, R4**) We replace Conv2d-based features with I3D features into an RNN model and obtain BLEU-4 of 11.27 and ROUGE-L of 32.22, similar to TSPNet-Single. Since our I3D backbone captures the spatial-temporal clues from video segments, it will leverage neighboring sharp frames to infer motion blurred sign gestures. In addition, our hierarchical feature learning effectively exploits windowing segments at different granularities, facilitating to capture fine-grained gestures. (ii) (**R3**) We clarify [2] uses AlexNet (see Sec. 5 in [2]), while Koller et al. and Cihan et al. use GoogleNet initialized on ILSVRC and finetuned on RWTH with glosses.

User studies (R3). Due to the diversity of translations, standard metrics may not fully reflect whether translation results convey the correct meanings or not. *To further demonstrate our performance gains*, we ask two participants to compare the predictions of TSPNet-Joint and [2] on RPWT to ground-truth translations, and choose the semantically more relevant result. The two participants favor 562 (87.54%) and 580 (90.34%) translations of TSPNet-Joint over those of [2] out of 642 testing instances. This experiment further demonstrates our significant performance improvements.

Difference between SLT and generic action localization (R3). Since signing progresses continuously, there is no clear boundary between consecutive sign gestures. Thus, it is ambiguous whether each sign gesture or an entire sign sequence should be treated as an action. Moreover, natural languages differ from sign languages grammatically. As a result, segmenting continuous signing into isolated signs is very difficult, especially in the absence of glosses. In contrast, as for the action localization tasks, temporal boundaries of an activity are evidently clear. This results in the essential difference between these two tasks. Although our approach is designed for SLT, we agree with R3 that it would be interesting to see our method generalizes to weakly supervised action localization, where only action labels are provided without boundary annotations. Due to the rebuttal timeframe, we have to leave this in future work.

Computation cost (R3). TSPNet-Joint and TSPNet-Single share the same architecture while TSPNet-Joint has 18% more operations for hierarchical feature learning over multi-granularity features. TSPNet-Joint takes 110 ms on average to process an RPWT video on an NVIDIA V100 GPU. We will include the discussion in the revision.

Segmenting isolated signs (R4). Please note that we did not claim our method achieves accurate temporal segmentation. Due to the grammatical differences between sign languages and natural languages, it is very difficult to segment sign gestures especially when glosses are unavailable. This motivates us to explore multi-scale windowing segments to learn representative features for sign language translation in an end-to-end fashion. Furthermore, Table 3 indicates that it is non-trivial to integrate multi-scale segment features. In this regard, our hierarchical feature learning method provides an effective solution to representing gestures for sentence translation in the absence of gloss annotations.