We thank all Reviewers for their constructive comments and insightful suggestions.

To Reviewer 1:
(1) Thank you for the very positive comments on our paper and kind suggestions about TRECVID dataset, we believe more impressive results can be provided in final version.

To Reviewer 2 & Reviewer 3:
(1) Regarding performance improvement, as shown in Table 1 in paper, the high-order pooling and the number of layers (6) Regarding the order (5) Regarding the depth, we test HPFN up to 4 layers (Table 2 in paper) and the optimal number of layers for IEMOCAP with Reviewer that it would be nice by adding more qualitative analysis in the final version.

To Reviewer 2:
(2) Time complexity: PTP is comparable or similar to LMF; HPFN is less than \( \sum_{l=1}^{L} N_l \) times of LMF, depending on specific architecture design choices such as number of layers, number of windows, window size and etc.

To Reviewer 3:
(1) Thank you for the very positive comments on our paper and kind suggestions about TRECVID dataset, we believe more impressive results can be provided in final version.
(2) Regarding training curves, we illustrate HPFN L1 & L2 and LMF in Figure in this page showing that HPFN is better than LMF. More comparisons on training curves will be added in the final version.
(3) It is a great suggestion to include significance testing, such as p-value test, in the paper.

Parameter Complexity: for PTP, due to the symmetry of feature tensor as well as the symmetric weight tensor, the number of parameters is independent of order \( P \) and linearly scales with the concatenated mixed features. For \( L \)-layer HPFN, number of parameters is linearly related to the number of ∑ segment=1^L N_l. In practice, \( N_1 \) is usually small and decreasing along layers, e.g., \( N_1 > N_2 > ... > N_L \). In our tests, complete/partial sharing strategy makes \( N_l \) even smaller. In principle, the parameter of HPFN is larger than or comparable to LMF (as HPFN is more powerful with temporal modelling), but significantly less than TFN. Please refer to the Table in this page.

The tradeoff is, if we employ more layers (or with more intermediate nodes in each layer) we get greater expressive capability. In practice, we need to choose optimal one so as not to overfit.

Table 1: \( I_f \) is output length. \( M \) is number of modalities. \( r \) is tensor rank. For PTP, [\( T, S \)] is the ‘local window size’ with \( S \leq M \). \( I_{r,m} \) is the dimension of features from modality \( m \) at time \( t \). For HPFN, \( N_l \) is the number of PTP ‘windows’ at layer \( l \in [L] \).