A Appendix

A.1 Reproducibility of the paper

We implement experiments on GTX 1080Ti. The main hyperparameters include audio hidden dimension, video hidden dimension, text hidden dimension, audio dropout rate, video dropout rate, text dropout rate, learning rate, weight decay, rank1 and rank2. Grid search is employed to find the appropriate combination of parameters. For each method, we randomly try 2000 combinations, since the model is small and the running time is short as shown in Table 1. The feature extraction method and the division of training and test sets follow (Zadeh et al., 2018). If the paper is accepted, we promise to open the source code and the best-performing hyperparameters.

Table 1: The model size and execution time of our methods.

Method	Size(MB)	Time(s)
TFN	1163	19.3
LMF	627	11.7
FT-LMF	1097	17.9
Dual-LMF	721	13.0

A.2 Derivations for Eqns. 5 and 6 in the paper

$W_h^i \cdot \tilde{V}$ can be rewritten as:

$$W_h^i \cdot \tilde{V} = \left[\sum_{r=1}^{R} \bigotimes_{m=1}^{M} (W_h^i)_{m,r}\right] \cdot \left[\bigotimes_{m=1}^{M} v_m\right] \tag{1}$$

where "$\cdot$" denotes linear operation for $\bigotimes_{m=1}^{M} v_m$. Since $\sum_{r=1}^{R} \bigotimes_{m=1}^{M} (W_h^i)_{m,r}$ and $\bigotimes_{m=1}^{M} v_m$ have the same size $R^{\prod_{m=1}^{M} d_m}$, we can rewrite the linear operation as the combination of element-wise multiplication and summation. The two formations are equivalent.

$$\begin{aligned} W_h^i \cdot \tilde{V} &= \sum \left[\sum_{r=1}^{R} \left[\bigotimes_{m=1}^{M} (W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m\right]\right] \\ &= \sum_{r=1}^{R} \left[\sum \left[\bigotimes_{m=1}^{M} (W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m\right]\right] \end{aligned} \tag{2}$$

where $\bigotimes_{m=1}^{M} (W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m$ can be rewritten as another formation, $\bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r} \circ v_m\right]$. The equivalence can be proven by element-wise comparison:

Proposition 1.

$$\bigotimes_{m=1}^{M} (W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m = \bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r} \circ v_m\right] \tag{3}$$

Proof.

$$\begin{aligned} &\left[\bigotimes_{m=1}^{M} (W_h^i)_{m,r} \circ \bigotimes_{m=1}^{M} v_m\right]_{c_1,c_2,\dots,c_M} \\ &= \left[\bigotimes_{m=1}^{M} (W_h^i)_{m,r}\right]_{c_1,c_2,\dots,c_M} \circ \left[\bigotimes_{m=1}^{M} v_m\right]_{c_1,c_2,\dots,c_M} \\ &= \left[(W_h^i)_{1,r}\right]_{c_1} \circ \dots \circ \left[(W_h^i)_{M,r}\right]_{c_M} \circ (v_1)_{c_1} \circ \dots \circ (v_M)_{c_M} \\ &= \left\{\left[(W_h^i)_{1,r}\right]_{c_1} \circ (v_1)_{c_1}\right\} \circ \dots \circ \left\{\left[(W_h^i)_{M,r}\right]_{c_M} \circ (v_M)_{c_M}\right\} \\ &= \left[(W_h^i)_{1,r} \circ v_1\right]_{c_1} \circ \dots \circ \left[(W_h^i)_{M,r} \circ v_M\right]_{c_M} \\ &= \left[\bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r} \circ v_m\right]\right]_{c_1,c_2,\dots,c_M} \end{aligned} \tag{4}$$

where $c_1, c_2, \dots, c_M (c_m \in [1, 2, \dots, d_m])$ denotes the index of the elements in high-order tensor. □

$W_h^i \cdot \tilde{V}$ can be rewritten as follows:

$$W_h^i \cdot \tilde{V} = \sum_{r=1}^{R} \left[\sum \bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r} \circ v_m\right]\right] \tag{5}$$

where $\sum \bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r} \circ v_m\right]$ can be rewritten as another formation, $\bigwedge_{m=1}^{M} \left[(W_h^i)_{m,r}^{\mathsf{T}} v_m\right]$. The equivalence can be proven as follows:

Proposition 2.

$$\sum \bigotimes_{m=1}^{M} \left[(W_h^i)_{m,r} \circ v_m\right] = \bigwedge_{m=1}^{M} \left[(W_h^i)_{m,r}^{\mathsf{T}} v_m\right] \tag{6}$$

Proof.

$$\bigwedge_{m=1}^{M} \left[(W_h^i)_{m,r}^{\mathsf{T}} v_m\right] = \bigwedge_{m=1}^{M} \sum \left[(W_h^i)_{m,r} \circ v_m\right] \tag{7}$$

Following the simple transformation like $(a+b)(c+d) = ac + ad + bc + bd$, we can easily transform $\bigwedge_{m=1}^{M} \sum \left[(W_h^i)_{m,r} \circ v_m\right]$ to

$\sum \bigotimes_{m=1}^{M}\left[(W_h^i)_{m,r} \circ v_m\right]$. These two formations are equal, just with different operation orders. The former utilizes summation($\sum$) first, while the later uses multiplication($\bigotimes$) between different elements first. $\square$

Therefore, we obtain the final formation of $W_h^i \tilde{V}$:

$$W_h^i \cdot \tilde{V} = \sum_{r=1}^{R} \bigwedge_{m=1}^{M}\left[(W_h^i)_{m,r}^{\mathsf{T}} v_m\right] \tag{8}$$

A.3 Derivations for Eqn. 17 in the paper

$W_k^j \cdot H_i$ can be rewritten as:

$$W_k^j \cdot H_i = \left[\sum_{r_2=1}^{R_2} \bigotimes_{m=1}^{M} (W_k^j)_{m,r_2}\right] \cdot \left[\sum_{r_1=1}^{R_1} \bigotimes_{m=1}^{M}\left[(W_h^i)_{m,r_1}^{\mathbf{T}} V_m\right]\right] \tag{9}$$

similar to Eqns. 2, 5, and 8, we obtain the final formation of $W_k^j \cdot H_i$,

$$\begin{aligned} & W_k^j \cdot H_i \\ &= \sum\left[\sum_{r_2=1}^{R_2}\sum_{r_1=1}^{R_1}\left[\bigotimes_{m=1}^{M}(W_k^j)_{m,r_2} \circ \bigotimes_{m=1}^{M}[(W_h^i)_{m,r_1}^{\mathbf{T}} V_m]\right]\right] \\ &= \sum\left[\sum_{r_2=1}^{R_2}\sum_{r_1=1}^{R_1}\bigotimes_{m=1}^{M}\left[(W_k^j)_{m,r_2} \circ [(W_h^i)_{m,r_1}^{\mathbf{T}} V_m]\right]\right] \\ &= \sum_{r_2=1}^{R_2}\sum_{r_1=1}^{R_1}\left[\sum \bigotimes_{m=1}^{M}\left[(W_k^j)_{m,r_2} \circ [(W_h^i)_{m,r_1}^{\mathbf{T}} V_m]\right]\right] \\ &= \sum_{r_2=1}^{R_2}\sum_{r_1=1}^{R_1}\bigwedge_{m=1}^{M}\left[(W_h^i)_{m,r_1}^{\mathbf{T}} V_m (W_k^j)_{m,r_2}\right] \end{aligned} \tag{10}$$

References

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *AAAI*.