Table 4: More ablation results for co-attention module.

	Livebot					VideoIC				
Variant	**R@1**	**R@5**	**R@10**	**MR**$^{\downarrow}$	**MRR**	**R@1**	**R@5**	**R@10**	**MR**$^{\downarrow}$	**MRR**
L=2	24.10	48.86	63.35	11.90	0.366	27.99	53.98	67.29	9.900	0.402
L=4	25.58	50.12	65.40	11.20	0.384	29.25	55.17	69.77	9.530	0.420
CoA1	24.73	50.33	64.95	11.19	0.379	27.33	53.29	68.15	10.22	0.402
CoA2	22.72	48.89	63.56	11.88	0.358	22.23	49.19	64.67	11.23	0.356
So-TVAE	**25.88**	**50.64**	**65.68**	**11.10**	**0.384**	**29.58**	**55.35**	**69.67**	**9.538**	**0.413**

A Co-attention Module

A.1 Definition

The co-attention module adopts L Transformer encoding layers follow by a weighting layer to achieve multi-modal interaction.

$$\hat{\mathbf{X}}, \hat{\mathbf{Y}} = \text{Co-Attention}(\mathbf{X}, \mathbf{Y}) \tag{19}$$

Taking X, Y as the input, and the detailed implementation is:

$$\mathbf{X}^l = \text{Transformer}_x^l(\mathbf{X}^{l-1}, \mathbf{X}^{l-1}, \mathbf{X}^{l-1}), \quad \mathbf{Y}^l = \text{Transformer}_y^l(\mathbf{Y}^{l-1}, \mathbf{X}^l, \mathbf{X}^l). \tag{20}$$

$$\hat{\mathbf{X}} = \sum_{i=1}^{k} \alpha_i X_i^L, \quad for \ \ \{\alpha_i\}_1^k = \text{softmax}(\text{MLP}(\mathbf{X}^L)), \tag{21}$$

where $\mathbf{X}^0 = \mathbf{X}$ and $\mathbf{Y}^0 = \mathbf{Y}$. $\hat{\mathbf{Y}}$ is obtained with a similar processing as Equation 21.

A.2 More Ablation Results for Co-attention

To further explore the structural rationality of co-attention module, we compared the following variants: (1) the models with different number L of Transformer encoding layers cascaded in depth. (2) CoA1: the model with co-attention acting on $(\mathbf{V_I}, \mathbf{V_e})$. (3) CoA2: the model with two co-attention acting on $(\mathbf{V_e}, \mathbf{V_I})$ and $(\mathbf{V_I}, \mathbf{V_e})$ respectively. The results shown in Table 4 verified the effectiveness of the co-attention module for the full model.

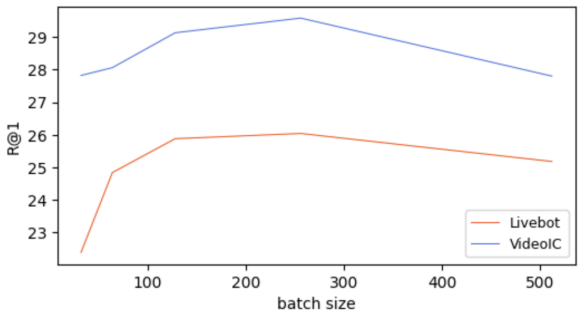

Figure 5: Comparison results based on batch size for batch attention module.

B Batch Attention Module

In this section, we further explored the impact of batch size on the sample interaction in a mini-batch. We compare the performance with the batch size set to 32, 64, 128, 256 and 512 respectively. From the comparison results shown in Figure 5, we can observe that the model achieves the best performance with the batch size is 256.