Table 2: Ablation study on the *TGIF-QA* dataset. For evaluation metric, Action, Trans. and FrameQA use ACC (%), while Count adopts Mean Square Error (MSE). V and T indicates visual attention and text attention. N indicates the number of video segments. ST is the best published methods using the same features as our model uses.

Method	Action	Trans	Frame	Count
STA-V-T(N=1)	71.94	78.67	56.01	4.26
STA-V-T(N=2)	72.16	78.84	55.92	4.25
STA-V-T(N=3)	**72.38**	78.96	56.52	4.27
STA-V-T(N=4)	72.34	**79.03**	**56.64**	**4.25**
STA-V(N=1)	71.02	77.24	54.89	4.32
STA-V(N=2)	71.46	77.47	55.05	4.37
STA-V(N=3)	71.68	77.31	55.50	4.34
STA-V(N=4)	71.24	77.57	55.47	4.33
ST (Jang et al. 2017)	59.04	65.56	45.60	4.55

FrameQA, Action and Trans., the accuracy (ACC.) is employed to evaluate model performance. Thus, the higher the ACC. is, the better the model is. The performance of Count is measured by Mean Square Error (MSE) between the true answer and the predicted answer. Therefore, lower MSE value indicates better performance of the model.

Implementation Details

For fair comparisons with recent work (Jang et al. 2017; Gao et al. 2018a), we use the same network ResNet152 (He et al. 2016) pre-trained on ImageNet 2012 classification to extract frame features. More specifically, all the frame features are obtained from the same pooling layer (pool5) and their dimension is 2048. In all tables in the experimental section, we use *R* to indicates that the input video's feature is extracted from ResNet152 feature. In addition, to reduce redundant information and reduce computation cost, we sample 36 frames from each video with equal spacing.

For text representation, we first encode each word with a pre-trained GloVe embedding to generate a 300-D vector following (Jang et al. 2017; Gao et al. 2018a). All the words are further encoded by a one-layer LSTM, whose hidden state has the size of 512. All the hidden states are concatenated and used for co-attention.

Training Details. In our experiments, the optimization algorithm is Adamax. The batch size is set as 128. The train epoch is set as 30. In addition, gradient clipping, weight normalization and dropout are employed in training. In addition, our implementation is based on the Pytorch library.

Ablation Study

The framework of our proposed STA consists of multiple major components. In order to evaluate the contribution of each component to the final performance, we conduct several ablation studies on the *TGIF-QA* dataset. Experimental results are shown in Tab. 2.

The Role of Structured Segmentation. The first block of Tab. 2 shows the effect of N, which is the number of structured segments. From the first block, we found that $N = 4$ improves performance of all four tasks to a certain extent. One possible reason is that dividing a video into multiple segments to conduct attention has the potential to locate the most relevant frames as well as to learn the long structures. However, the performance improvement with the increase in N is minor and the reason might be the nature of the GIF videos, which are well segmented and carefully curated, with an average length of 47 frames. Thus the advantages of our structured segment are not fully exploited.

Table 3: Comparison with the state-of-the-art method on the *TGIF-QA* dataset. R indicates ResNet152 features. For video representation, all methods take video spatial vectors only as the visual inputs.

Model	Action	Trans	Frame	Count
VIS+LSTM(agg)(R)	46.8	56.9	34.6	5.09
VIS+LSTM(avg)(R)	48.8	34.8	34.0	4.80
VQA-MCB(agg)(R)	58.9	34.3	24.7	5.17
VQA-MCB(avg)(R)	29.1	32.0	14.5	5.54
CT-SAN(R)	56.1	64.0	39.6	5.13
ST(R)	59.0	65.5	47.6	4.55
STA(R)	**72.3**	**79.0**	**56.6**	**4.25**

Table 4: Comparison with the state-of-the-art multi-feature based methods on the *TGIF-QA* dataset. R, C and F indicate Resnet152, C3D and Optical Flow features, respectively. Sp and Tp indicate spatial attention and temporal attention respectively.

Model	Action	Trans	Frame	Count
ST (R+C)	60.1	65.7	48.2	4.38
ST-Sp (R+C)	57.3	63.7	45.5	4.28
ST-Sp-Tp (R+C)	57.0	59.6	47.8	4.56
ST-Tp (R+C)	60.8	67.1	49.3	4.40
ST-Tp (R+F)	62.9	69.4	49.5	4.32
Co-memory (R+F)	68.2	74.3	51.5	**4.10**
STA (R)	**72.3**	**79.0**	**56.6**	4.25

The Role of Two-stream Attention. To analyze the contribution of two-stream attention: 1st-stream visual attention and 2nd-stream text attention, we conduct the second ablation study by removing the text attention and keeping the visual attention, which is represented as STA-V in Tab.2. From the table, we can see that with the same setting of N, STA-V-T performs better than STA-V. When $N = 4$, STA-V-T surpasses STA-V by 1.1% on Action, 1.46% on Trans. and 1.17% on FrameQA, and reduces MSE to 4.25 on Count. This ablation study shows the beneficial effects of our text attention.

In order to examine the influence of the visual attention, we compare our STA-V with the best published results obtained by the spatial-temporal reasoning (ST) method (Jang et al. 2017). From the last block we can see that our STA-V significantly outperforms ST by a large margin (12.24%, 12.07%, 9.87% on Action, Trans. and FrameQA, respectively). In addition, for Count, compared with ST with MSE of 4.55, STA-V-T reduces the error score to 4.33.

Qualitative Results

To understand the effect of our attention mechanism, we show some examples in Fig. 2. The first row demonstrates