Appendix

A. Motion vector warping accuracy

In this section, we verify the accuracy of warped features by motion vectors. We define the error rates e between the warped feature $W(f_s^t)$ by motion vector and the original feature f_s^{t+1} extracted from an actual P-frame using a pre-trained CNN as:

$$e = \frac{1}{n}\sum |f_s^{t+1} - W(f_s^t)| \times 100 \qquad (12)$$

where n is size of feature dimension.

Figure 1 illustrates comparison results between warped features and the original feature. The first example in Figure 1 shows the case when motion vectors are relatively accurate due to homogeneous motions of foreground objects. In this case, the error rate is 0.52 %, which shows quite high accuracy.

On the other hand, the second example in Figure 1 shows an opposite case when motion vectors are relatively inaccurate. In this case, the error rate is 12.47%. The residue increases with an incorrect motion vector. As shown in Eq. (7) of the paper, α obtained by residue reduces the influence of the attended motion features G_m by giving lower weight factors. In this manner, even though the proposed algorithm cannot use very precise motion vectors, it can tackle error propagation and generate the final appropriate feature by controlling α.

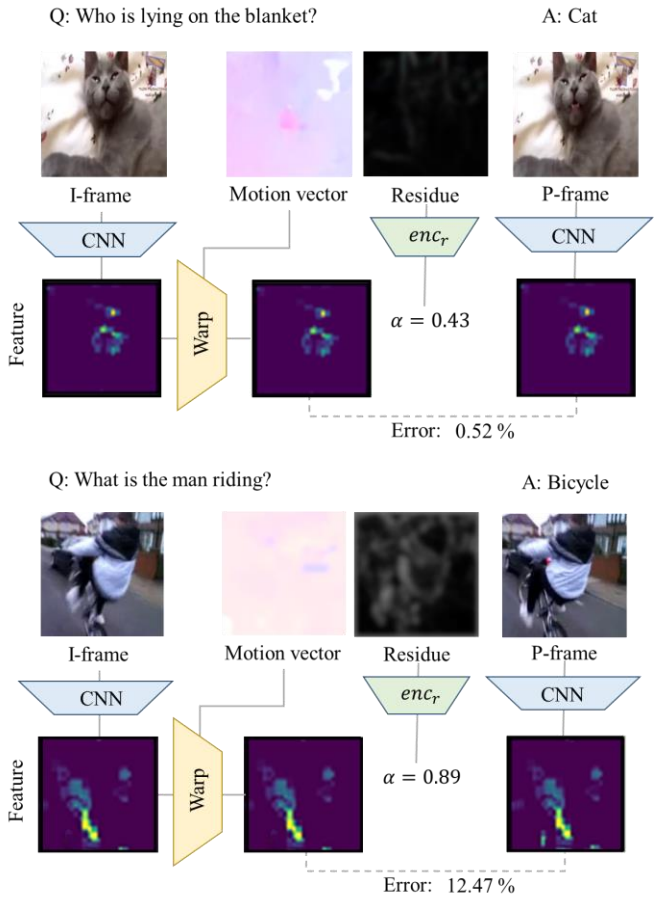

Figure 1 Visualization of motion vectors and features.

B. Comparison with algorithms of other tasks using compressed domain features

B.1 Action recognition task

We examine the performance of our compressed-domain features applied to an action recognition task. We show a comparative study with OF-CNN [R1], DTMV-CNN [R2], CoViAR [38], and DMC [29] in Table 1. The compared methods are developed for the action recognition although our features are originally developed for a video QA task. DTMV-CNN, CoViAR, and DMC use compressed-domain features as in the proposed technique. DTMV-CNN and CoViAR use refinement MVs instead of optical flow. DMC utilizes MVs and residue to increase an accuracy using discriminator. These all methods directly use MVs or rectified MV as an input of CNNs. OF-CNN applies two stream networks of which inputs are frame and optical flow.

To change the main task, VQAC(Base) removes the question attention modules in Eq. (3), (4), (5) and uses f_m^t and f_s^t instead of G_m^t and G_s^t, respectively, in Eq. (7). Moreover, we replace the fusion and decoder layers with an action classifier. We use the classifier of DMC (Resnet). The main difference with previous methods is that we use MVs to warp the I-frame feature and temporally combine features.

Table 2 Performance comparisons in HMDB51 [R3]

HMDB51	OF-CNN	DTMV-CNN	CoViAR	DMC	Our
Acc.	60.0	55.3	59.1	62.8	60.8

It is noted that, for this task, the question feature map and the question-guided attention have been disabled because our feature is not originally designed for action classification. Nevertheless, we achieve 60.8 % accuracy. The performance of the proposed algorithm is approximately the same as OF-CNN and approximately 2.0 % lower than DMC. The result implies that our compressed-domain video features exploit MVs and residues efficiently not only for the video QA task but also for motion-relevant CV tasks.

B.2 Video QA task

In this section, we compare the performance of an action-recognition network in video QA to consider the appropriate integrated networks. We integrate DMC [29] to HME, referred to as DMC(HME), by replacing a 3D-CNN feature extraction module with features from a MV classifier.

Table 2 Action -> QA on MSVD, MSR-VTT

	MSVD	MSR-VTT	Youtube2Text
HME	33.7	33.0	80.8
DMC (HME)	31.9	32.3	78.4
VQAC (HME)	37.8	35.7	82.1

DMC(HME) is even worse than HME about 1.8 %, 0.7 % and

5.8 % on MSVSD, MSR-VTT, and Youtube2Text, respectively. DMC was originally developed for speeding up a video action recognition. Therefore, the DMC motion feature is less suitable than the original motion feature from C3D on HME. Our integrated model outperforms in three datasets in Table 2. This experimental result demonstrates that our integrated network using the compression-domain features for a video QA task is designed properly. In contrast, the action recognition networks are designed to capture a dominant motion of a video. However, the experimental results demonstrated that such motion features are not suitable for the video QA task.

C. Qualitative performance evaluation

We conduct a bootstrap sampling five times and report means μ and standard deviations σ to show a fairly standard practice in Table 3.

Table 3 Results of a bootstrap sampling. Only AMU, COMEM, and HME provided available codes in the pampered methods.

MSVD dataset $\mu(\sigma)$				
AMU	VQAC(B)	COMEM	HME	VQAC(H)
31.4 (0.50)	31.0 (0.37)	31.3 (0.28)	32.9 (0.14)	37.3 (0.27)

We additionally evaluate our algorithm on Youtube2Text QA dataset [44] answering a multiple-choice problem to reveal different aspects of tested methods other than MSVD-QA and MSRVTT-QA.

Table 5 Visualization of motion vectors and features.

Method	Youtube2Text QA (Multi-choice)			
	What (2489)	Who (2004)	Others (97)	All (4590)
r-ANL	63.3	36.4	84.5	52.0
HME	83.1	77.8	88.6	80.8
VQAC(Base)	72.7	74.9	76.3	77.5
VQAC(HME)	79.1	85.3	90.3	82.1

The Table 4 presents the performance comparisons with r-ANL [44] and HME [12]. It is demonstrated that VQAC(Base) increases the accuracy of approximately 25.5% than r-AML. Moreover, VQAC(HME) provided a superior performance to HME.

We attempt to replace on VQAC(Base) with those from Resnet152, VGG16, and VGG19 and observe the performance is 31.5%, 31.4% and 31.5%, respectively. It displays that the model provides reliable performance with different spatial features.

We observe that overall performance is degraded when the I-period is larger (i.e. a smaller number of I-frames in a sequence and less) as shown in Table 6. The performance of dynamic videos is degraded more rapidly in both models, as expected.

Table 6 Results with various I-periods in MSVD QA. A video is categorized to a dynamic set (D) if noticeable motion changes are larger than 3. Otherwise, it is divided to a homogenous set (H).

I period	16	32	48	64
VQAC(Base) H/D	31.6 / 31.5	30.2 / 29.3	29.8 / 26.6	24.5 / 20.3
VQAC(HME) H/D	37.8 / 37.7	37.7 / 37.1	36.4 / 33.7	31.3 / 27.4

D. Quantitative performance evaluation

In this section, quantitative performance is further evaluated. We show exemplar cases where the scene of the video changes several times or the motion of the object in the video is large in Figure 3. For example, a video in the third row shows two people dancing violently on the grass. When the question “What are two people doing?” is given, previous algorithms predict the wrong answer, "play," but the proposed method predicts the right answer, "dance."

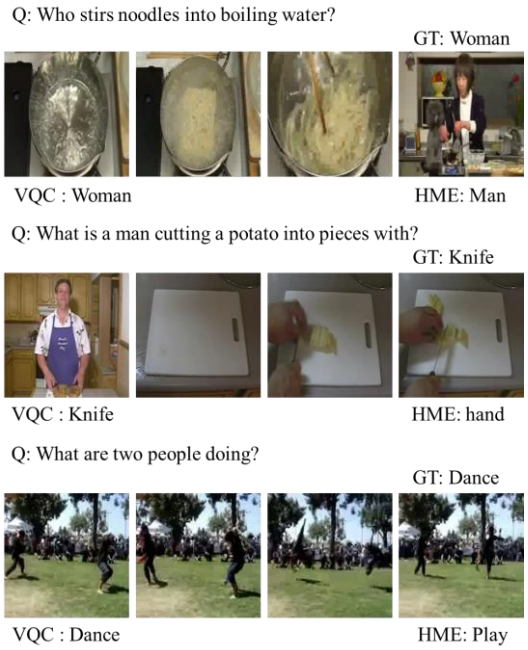

Figure 3 Qualitative performance evaluation in MSVD QA dataset.

E. Visualization of the question guided attention

The first and second example in Figure 4 illustrates that if the correct answer is about an object, the active area points the object correctly. For example, given the question "What is running around ballooning on floor?", the attention area is activated along with the “Dog”. The third example in Figure 4 shows that if the correct answer is about action, the attention area appears in an overall region of the image.

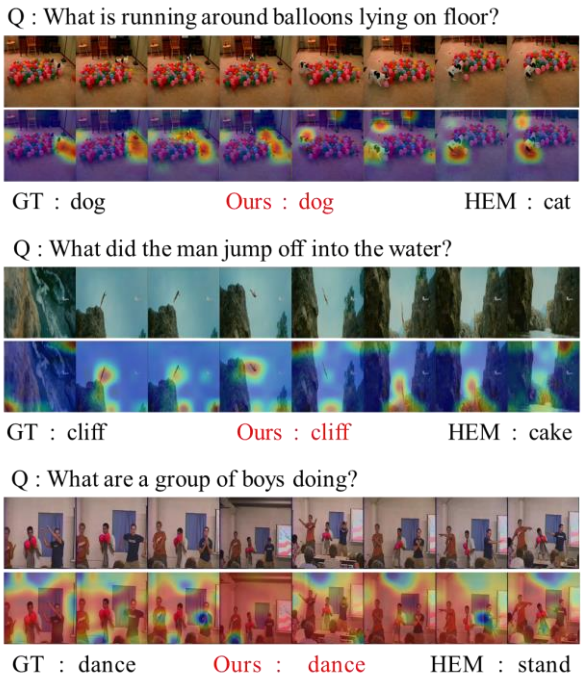

Figure 4 Visualization of the question guided attention $\boldsymbol{G_s^t}$ in MSVD QA dataset.

References

[R1] Simonyan and Zisserman. Two-stream convolutional networks for action recognition in videos. NIPS, Montreal, Canada, Dec. 2014, pp. 568–576.

[R2] Zhang, Bowen, et al. Real-time action recognition with deeply transferred motion vector cnns. IEEE Transactions on Image Processing, 2018, 27.5: 2326-2339.

[R3] Kuehne, Jhuang, Garrote, Poggio, and Serre. HMDB: A large video database for human motion recognition. IEEE Int. Conf. Comput. Vis. (ICCV), Nov. 2011, pp. 2556–2563