Audio-Visual Mismatch-Aware Video Retrieval via Association and Adjustment - *Supplementary Material* -

Sangmin Lee, Sungjune Park, and Yong Man Ro

Image and Video Systems Lab, KAIST, South Korea
{sangmin.lee, sungjune-p, ymro}@kaist.ac.kr

1 Network Structure Details

Table 1 shows the network structure details of the proposed model. The visual context embedder, visual semantic feed forward, associative feed forward, and mismatch-aware semantic embedder are included in MA-Transformer (See Figure 2 in the main manuscript). "Hidden Size" is the feature size after going through feed forward while "MLP Size" is the intermediate feature size of MHA. The projection head is a fc layer and is used in the associative learning procedure (See Figure 3 in the main manuscript). Further, the channel dimension d_v and d_a of memory m^v and m^a are 1024.

Table 1: Network structure details of the proposed model including layers in MA-Transformer and associative learning.

Network Structures				
Module	**Layers**	**Hidden Size**	**MLP Size**	**Multi-Heads**
Visual Context Embedder	3	1024	3072	8
Visual Semantic Feed Forward	1	1024	3072	-
Associative Feed Forward	1	1024	3072	-
Mismatch-Aware Semantic Embedder	3	1024	3072	8
Audio Semantic Embedder	1	1024	3072	8
Projection Head for Associative Learning	1	-	512	-

2 Video Retrieval Results

Table 2 shows the performance results of video retrieval on *MSR-VTT-1k-A*. As shown in the table, the proposed method also outperforms the other methods in terms of 1k-A split [60].

Table 2: Video to text retrieval performance comparison results on MSR-VTT according to data partition *MSR-VTT-1k-A*.

Method	Video Retrieval Performance				
	R@1↑(%)	R@5↑(%)	R@10↑(%)	MedR↓	mAP↑(%)
Masking Modalities [16]	22.5	53.2	67.1	4.7	-
Support-set Bottlenecks [45]	27.4	56.3	67.7	3.0	-
Proposed Method	**30.2**	**58.8**	**68.9**	**3.0**	**43.3**

3 Video to Text Retrieval Results

Table 3 shows the performance results of video to text retrieval on *MSR-VTT-Original*. The video to text retrieval is to find the corresponding text with a given video. As shown in the table, the proposed method also outperforms the other methods in terms of video to text retrieval, which indicates that video-text semantic matching is constructed properly.

Table 3: Video to text retrieval performance comparison results on MSR-VTT according to data partition *MSR-VTT-Original*.

Method	Text Retrieval Performance				
	R@1↑(%)	R@5↑(%)	R@10↑(%)	MedR↓	mAP↑(%)
W2VV [8]	17.0	37.9	49.1	11	7.6
VSE++ [13]	15.6	36.6	48.6	11	7.4
W2VV++ [31]	17.5	40.2	52.5	9	8.5
TCE [59]	15.1	36.8	50.2	10	8.0
HGR [5]	18.7	44.3	57.6	7	9.9
UWML [53]	18.2	45.2	58.7	-	-
HSL [10]	22.5	47.1	58.9	7	10.5
PSM [34]	22.8	48.0	61.0	6	11.6
T2VLAD [50]	20.7	48.9	62.1	6	-
Proposed Method	**23.9**	**51.3**	**64.3**	**5**	**12.9**

4 Effects of Memory Size

We perform experiments on the effects of the memory size k on video retrieval. The memory size k represents the number of slots in the memory. k is changed with an exponential scale (10, 50, 100, 500, and 1000) on VATEX dataset. As shown in Figure 1, mAP performance becomes saturated as k increases. This result shows the robustness to the setting of memory size.

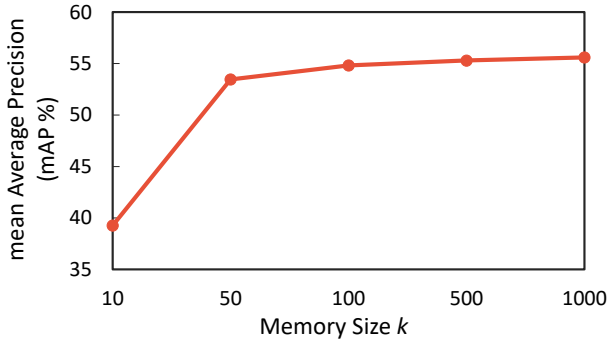

Fig. 1: Effects of the memory size k on the mAP performance of VATEX dataset.

5 Hyperparameters

Table 4, 5, 6 shows the performance changes according to temperatures (τ_m and τ_l), channel dimension (d_v and d_a), and loss margin (δ) parameters.

Table 4: Performance results according to τ_m and τ_l on VATEX.

$\tau_m \vert \tau_l$	0.1|0.1 ✓	1|0.1	0.5|0.1	0.05|0.1	0.1|1	0.1|0.5	0.1|0.05
mAP↑(%)	**55.3**	54.9	**55.3**	54.7	50.9	55.2	54.6

Table 5: Performance results according to d_v and d_a on VATEX.

d_v, d_a	256	512	1024 ✓	2048
mAP↑(%)	52.2	53.2	**55.3**	54.6

Table 6: Performance results according to δ on VATEX.

δ	0.1	0.2 ✓	0.3	0.4
mAP↑(%)	54.9	**55.3**	55.0	54.1

6 Qualitative Results

Figure 2 shows the qualitative results according to association and adjustment for mismatch cases. The first video does not include sound data and the retrieval result can be corrected by using the association. The second to fourth cases include mismatched audios and their visual semantics are distracted by the audios. In case of the third case, the audio includes noisy sound of the other instruments. Thus, the model without adjustment captures the wrong video with organ-like sound. The adjustment makes the model find the right video as shown.

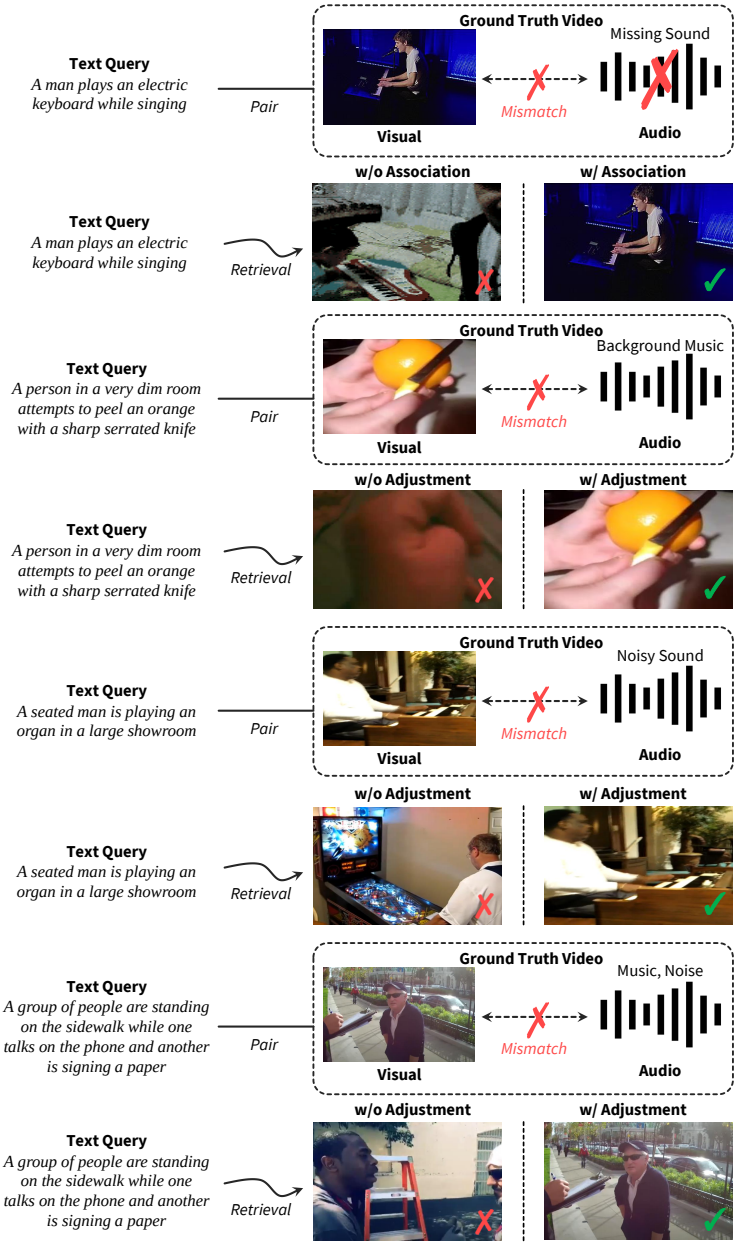

Fig. 2: Qualitative retrieval results for audio-visual mismatch cases.