Supplementary material of Lightweight Attentional Feature Fusion: A New Baseline for Text-to-Video Retrieval

Fan Hu[1,2], Aozhu Chen[1,2], Ziyue Wang[1,2], Fangming Zhou[1,2], Jianfeng Dong[3], and Xirong Li[1,2]

[1] MoE Key Lab of DEKE, Renmin University of China
[2] AIMC Lab, School of Information, Renmin University of China
[3] College of Computer and Information Engineering, Zhejiang Gongshang University

In this supplement, we provide more experimental results that are not included in the paper due to space limit.

Distribution of attentional weights per feature. We analyze the attentional weights per feature on different datasets, with mean and std values shown in Tab. 9. For both text and video features, *clip-ft* is predominant. Meanwhile, the feature-specific weights are relatively stable across datasets, suggesting that feature fusion patterns found by LAFF are also stable.

Table 9: **Mean and std of attentional weights per feature**.

Dataset	**Text features**				**Video features**			
	gru	*bow*	*w2v*	*clip-ft*	*tf*	*x3d*	*ircsn*	*clip-ft*
MV-test3k	0.10±0.06	0.16±0.12	0.06±0.04	0.68±0.15	0.15±0.15	0.12±0.12	0.11±0.11	0.62±0.62
MV-test1k	0.10±0.06	0.15±0.11	0.06±0.03	0.69±0.14	0.15±0.04	0.12±0.02	0.10±0.02	0.62±0.07
MSVD	0.13±0.07	0.16±0.07	0.12±0.06	0.58±0.10	0.15±0.05	0.15±0.04	0.07±0.02	0.63±0.07
TGIF	0.11±0.06	0.15±0.10	0.07±0.03	0.66±0.12	0.15±0.03	0.17±0.04	0.17±0.03	0.50±0.05
VATEX	0.09±0.03	0.15±0.07	0.07±0.02	0.68±0.08	0.16±0.03	0.17±0.04	0.18±0.04	0.48±0.05

How the individual embedding spaces differ from each other? To reveal how different are the embedding spaces to each other, we compute the Jaccard index between the top-5 video retrieval results of two spaces *w.r.t.* a specific query caption. As Tab. 10 shows, the inter-space Jaccard index is lower than 0.5, suggesting sufficient divergence.

Comparing different feature fusion blocks. A table similar to Tab. 4 but on the video features is shown in Tab. 11. Again, LAFF performs the best, followed by Attention-free, the concatenation baseline and MHSA.

Comparison with SOTA. The performance of the SOTA methods on MV-test3k, MV-test1k, MSVD, TGIF and VATEX is summarized in Tab. 12. Their *Med r* scores reported in Tab. 13, smaller is better.

Per-query analysis on TV20. Tab. 14 shows the performance of the individual test queries of TV20. The mean infAP scores of CLIP-FT and CLIP2Video

Table 10: **Inter-space Jaccard index**. Data: MV-test3k. Model: LAFF($h = 8$)

Space index	0	1	2	3	4	5	6	7
0	1.000	0.443	0.453	0.446	0.437	0.441	0.448	0.440
1	0.443	1.000	0.453	0.445	0.446	0.449	0.450	0.445
2	0.453	0.453	1.000	0.458	0.451	0.455	0.459	0.445
3	0.446	0.445	0.458	1.000	0.447	0.454	0.456	0.439
4	0.437	0.446	0.451	0.447	1.000	0.454	0.449	0.438
5	0.441	0.449	0.455	0.454	0.454	1.000	0.450	0.442
6	0.448	0.450	0.459	0.456	0.449	0.450	1.000	0.443
7	0.440	0.445	0.445	0.439	0.438	0.442	0.443	1.000

Table 11: **Comparing feature fusion blocks**. The simple feature concatenation used by W2VV++ is taken as a baseline. Numbers in parentheses are relative improvements against this baseline. Text features: $\{$*bow*, *w2v*, *gru*, *clip*$\}$. Data: MV-test3k.

Video features	**Fusion block**	**R1**	**R5**	**R10**	**Medr**	**mAP**
rx101, re152	Baseline	14.4	34.9	46.3	13	0.247
	MHSA	12.6	32.2	42.7	16	0.224 (9.31%↓)
	Attention-free	14.8	35.6	46.9	12	0.252 (2.02%↑)
	LAFF	**16.0**	**36.9**	**48.5**	**11**	**0.265 (7.29%↑)**
rx101, re152, wsl	Baseline	16.7	39.0	50.7	10	0.276
	MHSA	14.5	35.7	46.9	13	0.249 (9.78%↓)
	Attention-free	17.2	39.5	51.2	10	0.282 (2.17%↑)
	LAFF	**18.6**	**41.3**	**52.8**	**9**	**0.298 (7.97%↑)**
rx101,re152,wsl,clip	Baseline	17.8	41.1	52.7	9	0.291
	MHSA	19.1	42.9	54.2	8	0.306 (5.15%↑)
	Attention-free	19.4	43.3	54.8	8	0.310 (6.53%↑)
	LAFF	**23.7**	**48.5**	**59.8**	**6**	**0.356 (22.34%↑)**

are 0.172 and 0.180, respectively, 21.0% and 17.2% lower than LAFF. In particular, we notice that LAFF is better than CLIP-FT and CLIP2Video for answering action-related queries. The result suggests that the image features alone are insufficient. Visual features that capture motion/temporal information, *e.g. ircsn* and *c3d*, are necessary for video retrieval on TRECVID-like collections that contain quite diverse video content.

Table 12: **Comparison with SOTA**.

Model	MV-test3k			MV-test1k			MSVD			TGIF			VATEX		
	R1	*R5*	*R10*	*R1*	*R5*	*R10*	*R1*	*R5*	*R10*	*R1*	*R5*	*R10*	*R1*	*R5*	*R10*
JE, IJMIR19 [12]	7.0	20.9	29.7	n.a.	n.a.	n.a.	20.2	47.5	60.7	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
W2VV++, MM19 [9]	11.1	29.6	40.5	18.9	45.3	57.5	22.4	51.6	64.8	9.4	22.3	29.8	34.3	73.6	83.7
PIE-Net, CVPR19 [15]	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	3.0	9.7	14.9	n.a.	n.a.	n.a.
CE, BMVC19 [11]	10.0	29.0	41.2	20.9	48.8	62.4	19.8	49.0	63.8	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
TCE, SIGIR20 [16]	7.7	22.5	32.1	16.1	38.0	51.5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
HGR, CVPR20 [4]	9.2	26.2	36.5	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	4.5	12.4	17.8	35.1	73.5	83.5
SEA, TMM21 [10]	13.1	33.4	45.0	23.8	50.3	63.8	23.9	53.9	67.3	11.1	25.2	32.8	35.5	74.7	85.4
MMT, ECCV20 [8]	n.a.	n.a.	n.a.	26.6	57.1	69.6	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
DE, TPAMI21 [6]	11.6	30.3	41.3	21.1	48.7	60.2	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	36.8	67.5	78.9
Frozen, ICCV21 [2]	n.a.	n.a.	n.a.	31.0	59.5	70.5	33.7	64.7	76.3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
TEACHTEXT, ICCV21 [5]	15.0	38.5	51.7	29.6	61.6	74.2	25.4	56.9	71.3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
SSB, ICLR21 [13]	n.a.	n.a.	n.a.	30.1	58.5	69.3	28.4	60.0	72.9	n.a.	n.a.	n.a.	45.9	82.4	90.4
SSML, AAAI21 [1]	17.4	41.6	53.6	n.a.	n.a.	n.a.	20.3	49.0	63.3	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
CLIP, MCPR21 [14]	21.4	41.1	50.4	31.2	53.7	64.2	37.0	64.1	73.8	n.a.	n.a.	n.a.	39.7	72.3	82.2
CLIP-FRL, ICCVW21 [3]	22.9	47.0	57.9	38.2	66.0	75.7	33.9	64.9	76.3	13.2	29.2	37.9	47.1	82.3	90.6
CLIP-FT (*this paper*)	27.7	53.0	64.2	39.7	67.8	78.4	44.6	74.7	84.1	21.5	40.6	49.9	53.3	87.5	94.0
The same video and text feature as ours															
JE [12] (uniform weights)	21.2	46.5	58.4	36.0	65.9	76.4	35.9	71.0	81.8	18.7	37.5	47.1	50.2	88.7	95.4
JE (0.8 for *clip-ft*)	26.1	51.7	63.3	41.2	73.2	82.5	39.4	69.9	79.4	21.7	41.3	50.9	54.1	89.0	95.0
JE (0.9 for *clip-ft*)	25.9	51.4	63.0	40.9	72.7	82.1	38.8	69.7	78.9	21.3	40.9	50.3	53.5	88.3	94.6
W2VV++ [9]	23.0	49.0	60.7	39.4	68.1	78.1	37.8	71.0	81.6	22.0	42.8	52.7	55.8	91.2	96.0
SEA [10]	19.9	44.3	56.5	37.2	67.1	78.3	34.5	68.8	80.5	16.4	33.6	42.5	52.4	90.2	95.9
MMT [8]	24.9	50.5	62.0	39.5	68.3	78.3	40.6	72.0	81.7	22.1	42.2	51.7	54.4	89.2	95.0
LAFF	28.0	53.8	64.9	42.2	70.7	**81.2**	45.2	75.8	84.3	24.1	44.7	54.3	57.7	91.3	95.9
LAFF-ml	**29.1**	**54.9**	**65.8**	**42.6**	**71.8**	81.0	**45.4**	**76.0**	**84.6**	**24.5**	**45.0**	**54.5**	**59.1**	**91.7**	**96.3**
Comparison with arXiv SOTA															
CLIP2Video [7]	n.a	n.a	n.a	44.5	71.3	80.6	44.7	74.8	83.7	n.a	n.a	n.a	54.8	89.1	95.1
LAFF	n.a	n.a	n.a	**45.8**	**71.5**	**82.0**	**45.4**	**75.5**	**84.1**	n.a	n.a	n.a	**58.3**	**91.7**	**96.3**

Table 13: **Comparison with SOTA**. Performance metric: $Med\ r$.

Model	MV-test3k	MV-test1k	MSVD	TGIF	VATEX
JE, IJMIR19 [12]	34	n.a.	6	n.a.	n.a.
W2VV++, MM19 [9]	18	8	5	48	2
PIE-Net, CVPR19 [15]	n.a.	n.a.	n.a.	155	n.a.
CE, BMVC19 [11]	16	6	6	n.a.	n.a.
TCE, SIGIR20 [16]	30	10	n.a.	n.a.	n.a.
HGR, CVPR20 [4]	24	n.a.	n.a.	160	2
SEA, TMM21 [10]	14	5	5	35	2
MMT, ECCV20 [8]	n.a.	4	n.a.	n.a.	n.a.
DE, TPAMI21 [6]	16	n.a.	n.a.	n.a.	3
Frozen, ICCV21 [2]	n.a	3	3	n.a.	n.a.
TEACHTEXT, ICCV21 [5]	10	3	4	n.a.	n.a.
SSB, ICLR21 [13]	n.a.	3	4	n.a.	**1**
SSML, AAAI21 [1]	8	n.a.	6	n.a.	n.a.
CLIP, MCPR21 [14]	10	4	3	n.a.	2
CLIP-FRL, ICCVW21 [3]	7	**2**	3	25	2
CLIP-FT (*this paper*)	5	**2**	**2**	11	**1**
The same video and text feature as ours					
JE (uniform weights)	7	3	**2**	13	**1**
JE (0.8 for *clip-ft*)	5	**2**	**2**	10	**1**
JE (0.9 for *clip-ft*)	5	**2**	**2**	10	**1**
W2VV++	6	**2**	**2**	9	**1**
SEA	7	**2**	3	17	**1**
MMT	5	**2**	**2**	9	**1**
LAFF	**4**	**2**	**2**	**8**	**1**
LAFF-ml	**4**	**2**	**2**	**8**	**1**
Comparison with the SOTA on arxiv					
CLIP2Video	n.a	**2**	**2**	n.a	**1**
LAFF	n.a	**2**	**2**	n.a	**1**

Table 14: **Performance of LAFF, CLIP-FT and CLIP2Video on the TV20 AVS task**. Colored numbers indicate over 0.1 absolute difference.

Query	**Object**	**Person**	**Action**	**Location**	**LAFF**	**CLIP-FT**	**CLIP2Video**
641 showing an aerial view of buildings near water in the daytime	✓			✓	0.297	0.240	0.234
642 a person paddling kayak in the water	✓	✓	✓	✓	0.399	0.445	0.430
643 people dancing or singing while wearing costumes outdoors		✓	✓		0.080	0.071	0.134
644 sailboats in the water	✓			✓	0.649	0.643	0.352
645 a person wearing a necklace	✓	✓			0.036	0.088	0.034
646 a woman sitting on the floor	✓	✓			0.091	0.106	0.163
647 people or cars moving on a dirt road	✓	✓	✓		0.199	0.223	0.267
648 a man in blue jeans outdoors	✓	✓		✓	0.018	0.081	0.053
649 someone jumping while snowboarding		✓	✓		0.448	0.451	0.785
650 one or more people drinking wine		✓	✓		0.054	0.040	0.096
651 one or more people skydiving		✓	✓		0.352	0.427	0.489
652 a little boy smiling		✓	✓		0.197	0.203	0.165
653 group of people clapping		✓	✓		0.082	0.120	0.349
654 one or more persons exercising in a gym		✓	✓	✓	0.227	0.179	0.253
655 one or more persons standing in a body of water		✓	✓	✓	0.018	0.031	0.023
656 a long haired man	✓	✓			0.277	0.217	0.400
657 a woman with short hair indoors	✓	✓		✓	0.061	0.054	0.020
658 two or more people under a tree	✓	✓		✓	0.035	0.058	0.034
659 a church from the inside	✓			✓	0.188	0.255	0.342
660 train tracks during the daytime	✓				0.361	0.211	0.524

References

1. Amrani, E., Ben-Ari, R., Rotman, D., Bronstein, A.: Noise estimation using density estimation for self-supervised multimodal learning. In: AAAI (2021)
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV (2021)
3. Chen, A., Hu, F., Wang, Z., Zhou, F., Li, X.: What matters for ad-hoc video search? a large-scale evaluation on TRECVID. In: ICCV Workshop on ViRal (2021)
4. Chen, S., Zhao, Y., Jin, Q., Wu, Q.: Fine-grained video-text retrieval with hierarchical graph reasoning. In: CVPR (2020)
5. Croitoru, I., Bogolin, S.V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., Liu, Y.: TEACHTEXT: Crossmodal generalized distillation for text-video retrieval. In: ICCV (2021)
6. Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X.: Dual encoding for video retrieval by text. TPAMI (2021)
7. Fang, H., Xiong, P., Xu, L., Chen, Y.: CLIP2Video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
8. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: ECCV (2020)
9. Li, X., Xu, C., Yang, G., Chen, Z., Dong, J.: W2VV++: Fully deep learning for ad-hoc video search. In: ACMMM (2019)
10. Li, X., Zhou, F., Xu, C., Ji, J., Yang, G.: SEA: Sentence encoder assembly for video retrieval by textual queries. TMM pp. 4351–4362 (2020)
11. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. In: BMVC (2019)
12. Mithun, N.C., Li, J., Metze, F., Roy-Chowdhury, A.K.: Joint embeddings with multimodal cues for video-text retrieval. IJMIR pp. 3–18 (2019)
13. Patrick, M., Huang, P., Asano, Y.M., Metze, F., Hauptmann, A.G., Henriques, J.F., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. In: ICLR (2021)
14. Portillo-Quintero, J.A., Ortiz-Bayliss, J.C., Terashima-Marín, H.: A straightforward framework for video retrieval using CLIP. In: MCPR (2021)
15. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval. In: CVPR (2019)
16. Yang, X., Dong, J., Cao, Y., Wang, X., Wang, M., Chua, T.S.: Tree-augmented cross-modal encoding for complex-query video retrieval. In: SIGIR (2020)