Renmin University of China and Tencent at TRECVID 2023: Harnessing Pre-trained Models for Ad-hoc Video Search

Xirong Li[1,2], Fan Hu[1,2], Ruixiang Zhao[1,2], Ziyuan Wang[1,2], Jingyu Liu[1,2], Jiazhen Liu[1,2], Bangxiang Lan[1,2], Wenguan Kou[1,2], Yuhan Fu[1,2], Zhanhui Kang[3]
[1]MOE Key Lab of DEKE, Renmin University of China
[2]AIMC Lab, School of Information, Renmin University of China
[3]Tencent
https://ruc-aimc-lab.github.io

Abstract

We summarize our TRECVID 2023 Ad-hoc Video Search (AVS) experiments. We focus on leveraging pre-trained multimodal models for video and text representation. For video feature extraction, we utilized pre-trained models including BLIP, CLIP, irCSN, BEiT, WSL, Video-LLaMA, and BLIP-2. For text features, we employed BLIP, CLIP, and BLIP-2. Our approach to feature fusion is based on the Lightweight Attentional Feature Fusion (LAFF) network, which has been used in our 2022 system. LAFF performs feature fusion at both early and late stages, and at both text and video ends. In addition to LAFF, we also tried TeachCLIP, a new text-to-video retrieval model developed jointly by AIMC-Lab@RUC and Tencent. TeachCLIP leverages multi-grained knowledge distillation to let a CLIP4Clip-based student network to learn from more advanced, yet computationally heavy models. Our training data inlcudes three manually annotated datasets (MSR-VTT, TGIF and VATEX), two webly annotated datasets (WebVid and ChinaOpen), and an auto-generated video description dataset (V3C1-PC). We trained 42 LAFFs with varied setups and one TeachCLIP in total. These 43 models are late fused in a learning-to-rank manner. The 2023 edition of the TRECVID benchmark has been a fruitful participation for the RUCMM-Tencent team. Our best run, with an infAP of 0.272, is ranked at the third place teamwise.

1 Our Approach

Our solution for the TRECVID 2023 AVS task focuses on harnessing pretrained large multimodal models for extracting diverse features from videos and textual queries. The varied features are combined by the Lightweight Attentional Feature Fusion (LAFF) network [8], which has been used with success in our TRECVID 2022 system [11]. LAFF performs feature fusion at both early and late stages and at both text and video ends. We also tried TeachCLIP [21], a new text-to-video retrieval model developed jointly by AIMC-Lab@RUC and Tencent. TeachCLIP leverages multi-grained knowledge distillation to let a CLIP4Clip [17] based student network learn from more advanced yet computationally heavy models. An overview of this year's video search engine is illustrated in Fig. 1.

1.1 LAFF-based Video Retrieval

Compared to our TRECVID 2022 system, we keep the framework of LAFF unchanged, but explore more recent pre-trained models to extract visual and textual features.

1.1.1 Choice of Visual Features

The following seven deep *visual features* are used:

1. *wsl*: A 2,048-d frame-level feature, extracted by ResNeXt-101 which pre-trained on weakly labeled web images followed by fine-tuning on ImageNet[1] [19].
2. *beit*: A 1,024-d frame-level feature, extracted by BEiT which pre-trained on full ImageNet and fine-tune on 1k-class ImageNet [2] [4].
3. *clip*: A 768-d frame-level feature, extracted by a pre-trained CLIP (ViT-L/14)@336 model. [3] [20].
4. *blip*: A 256-d frame-level feature, extracted by a BLIP(ViT-B) [10] which pre-trained on 129M image-text pairs [4].
5. *blip2*: A 256-d frame-level feature, extracted by a BLIP-2(ViT-G) [9] which pre-trained on 129M image-text pairs [5].
6. *ircsn*: A 2,048-d segment-level feature, extracted by irCSN-152 [7] which trained on IG-65M[6].

[1]https://github.com/facebookresearch/WSL-Images
[2]https://github.com/microsoft/unilm/tree/master/beit
[3]https://github.com/openai/CLIP
[4]https://github.com/salesforce/BLIP
[5]https://github.com/salesforce/LAVIS/tree/main/projects/blip2
[6]https://github.com/facebookresearch/VMZ/tree/master/pt

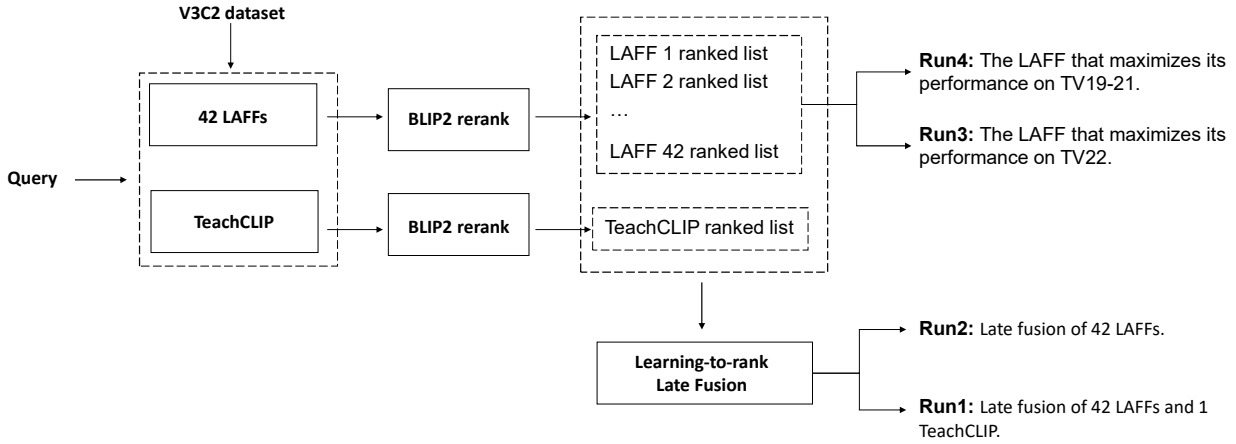

Figure 1: Overview of RUCMM Video Search Engine@TV23.

7. *video-llama*: A 768-d segment-level feature, extracted by Audio-Visual Language Model Video-LLaMA[7] [24].

1.1.2 Choice of Textual Features

We experimented with the following three sentence encoders for textual features:

1. Text encoder of CLIP(ViT-L/14)@336 [20], *i.e.* a GPT.
2. Text encoder of BLIP(ViT-B) [10],*i.e.* a BERT.
3. Text encoder of BLIP-2(ViT-G) [9], *i.e.* a BERT.

1.1.3 Choice of (Pre-)Training Data

Besides V3C1-PC used in our previous system [11], this year we add two webly annotated datasets, *i.e.* WebVid [3] and ChinaOpen [5] for pre-training, see Table 1.

- **WebVid2.5M**. We adopt the 2.5M version of WebVid [3]. Due to broken links and removal of over-length videos, we obtain around 2.3M web videos.
- **ChinaOpen**. ChinaOpen is a new video dataset targeted at open-world multimodal learning, with raw data gathered from Bilibili, a popular Chinese video-sharing website [5]. For each training video in ChinaOpen, we generate an English caption using BLIP2 for the middle frame, and use the generated caption as the video annotation.

As for training data, we use the joint set of MSR-VTT [23], TGIF [16] and VATEX [22] (M+T+V). Following our conventional setup [12–15], the development set of the TRECVID 2016 Video-to-Text Matching task [2] is used as an external validation set[8] for base model selection.

[7] https://github.com/DAMO-NLP-SG/Video-LLaMA
[8] https://github.com/li-xirong/avs

Table 1: Statistics of three pre-training datasets used in our TV23 AVS solutions.

Dataset	Frames	Video clips	Sentences
V3C1-PC [11]	219,531	1,605,335	436,204
ChinaOpen [5]	52,170	1,516,598	52,170
WebVid [3]	2,291,129	44,847,987	2,291,129

1.2 TeachCLIP for end-to-end Video Retrieval

For large scale text-to-video retrieval, efficiency is a factor that must be considered. To achieve a good balance between efficiency and effectiveness, we have developed in [21] TeachCLIP, which performs multi-grained knowledge distillation to let a CLIP4Clip based student network learn from more advanced yet computationally heavy models. To improve the student's learning capability, we add an Attentional frame-Feature Aggregation (AFA) block, which by design adds no extra storage / computation overhead at the retrieval stage. While attentive weights produced by AFA are commonly used for combining frame-level features, we propose a novel use of the weights to let them imitate frame-text relevance estimated by the teacher network. As such, AFA provides a fine-grained learning (teaching) channel for the student (teacher). We instantiate TeachCLIP's teacher model with X-CLIP [18], which has been pre-trained on the M+T+V dataset.

1.3 Search Result Reranking

Given a top-ranked list of (5k) videos returned by a base model, we re-score each video in the list by considering a fine-grained cross-modal similairty between its n frames and the given query. The frame-query similarity is computed based on their embeddings obtained by BLIP-2(ViT-G). Max pooling is used to aggregated the frame-level similarities to the video level. The new relevance score is obtained by a weighted linear fusion of the newly computed score (0.6) and the original score (0.4). See [6] for more details.

1.4 Late Fusion Method

In order to overcome the limitations of single model performance and improve generalization ability, we have trained 42 models based on LAFF and 1 model based on TeachCLIP with different combinations of features and training datasets, then we fused these 43 models with learned weight with a *two-step* learning-to-rank strategy.

In Step 1, we optimize the fusion weights by minimizing a ranking loss on the TV22 query set. The Adam optimizer is used, with the initial learning rate of 0.2, weight_decay of 1e-3 and 100 training epochs. Notice however that weights learned in a such manner can be suboptimal. So in Step 2, we employ a greedy search strategy, given the intial weights obtained in Step 1. In particular, we search each model weight sequentially from 0 to 1 with a stride of 0.1 to maximize the TV22 infAP score.

2 Internal Evaluation

2.1 Datasets for Pre-training

To assess the influence of different pre-training datasets for LAFF, we provide the compared results on TV19-22. we use all features for feature fusion and MSR-VTT+TGIF+VATEX for training. As Table 2 shows, V3C1-PC for pre-training has the best performance on TV22, and WebVid+V3C1-PC for pre-training has the best performance on MEAN(TV19-21).

With various training strategies and feature combinations, we got a total of 44 LAFF models.

Table 2: Evaluating the influence of different pre-training datasets on the TRECVID 19-22 AVS tasks.

Pre-training	TV19	TV20	TV21	MEAN(TV19-21)	TV22
V3C1-PC	0.255	0.345	0.352	0.317	**0.258**
WebVid	0.235	0.333	0.317	0.295	0.230
ChinaOpen	0.252	0.337	0.335	0.308	0.252
ChinaOpen, V3C1-PC	0.252	0.341	0.347	0.313	0.251
WebVid, V3C1-PC	0.256	0.349	0.351	**0.319**	0.251
ChinaOpen, WebVid, V3C1-PC	0.248	0.337	0.337	0.307	0.247

2.2 TeachCLIP for AVS

For TeachCLIP, we first pre-train X-CLIP on a dataset composed of three datasets: TGIF, MSRVTT, and VATEX. Then, we use it as a teacher for distillation, employ a trained student model for retrieval, and consider the obtained results as one of the models for later fusion. The performance of TeachCLIP is shown in Table 3. As TeachCLIP is inferior to LAFF, we do not use it alone. Rather, we fuse it with the LAFFs.

Table 3: Performance of TeachCLIP on the TRECVID 19-22 AVS tasks.

Model	TV19	TV20	TV21	MEAN(TV19-21)	TV22
TeachCLIP	0.115	0.149	0.166	0.143	0.134

2.3 Learning based Model Fusion

The result of the two-step strategy is shown in Table 4. Compared to the single best model, combining the multiple models with the Step 1 weights show a clear improvement. Greedy search in Step 2 further improves the performance, though marginally.

Table 4: Evaluating late fusion on the TV22 query set.

Fusion Models	Best Model	Step 1	Step 1+Step 2
42 LAFF	0.258	0.276	0.279
42 LAFF + 1 TeachCLIP	0.258	0.278	0.282

3 Submissions

We submitted four runs as follows:

- *Run 4*: LAFF that maximizes infAP on TV19-21, *i.e.* with BLIP, CLIP and BLIP-2 as its text features, *blip*, *clip*, *ircsn*, *beit*, *wsl*, *video-llama* and *blip2* as video features, pre-trained on WebVid, V3C1-PC.
- *Run 3*: LAFF that maximizes infAP on TV22, *i.e.* with BLIP and CLIP as its text features and *blip*, *clip*, *ircsn*, *beit*, *wsl* and *video-llama* as video features, pre-trained on V3C1-PC.
- *Run 2*: An ensemble of 42 LAFF models.
- *Run 1*: An ensemble of 42 LAFFs plus a TeachCLIP.

The performance of our runs on the TV23 AVS task is shown in Fig. 2. Our best run is Run 1, with a mean infAP of 0.272.

Acknowledgments

The authors are grateful to the TRECVID coordinators for the benchmark organization effort [1]. This research was supported by the National Natural Science Foundation of China (No. 62172420) and Tencent Marketing Solution Rhino-Bird Focused Research Program.

References

[1] G. Awad, K. Curtis, A. A. Butt, J. Fiscus, A. Godil, Y. Lee, A. Delgado, E. Godard, L. Diduch, D. Gupta, D. D. Fushman, Y. Graham, and G. Quénot. TRECVID 2023 – a series of evaluation tracks in video understanding. In *TRECVID*, 2023.

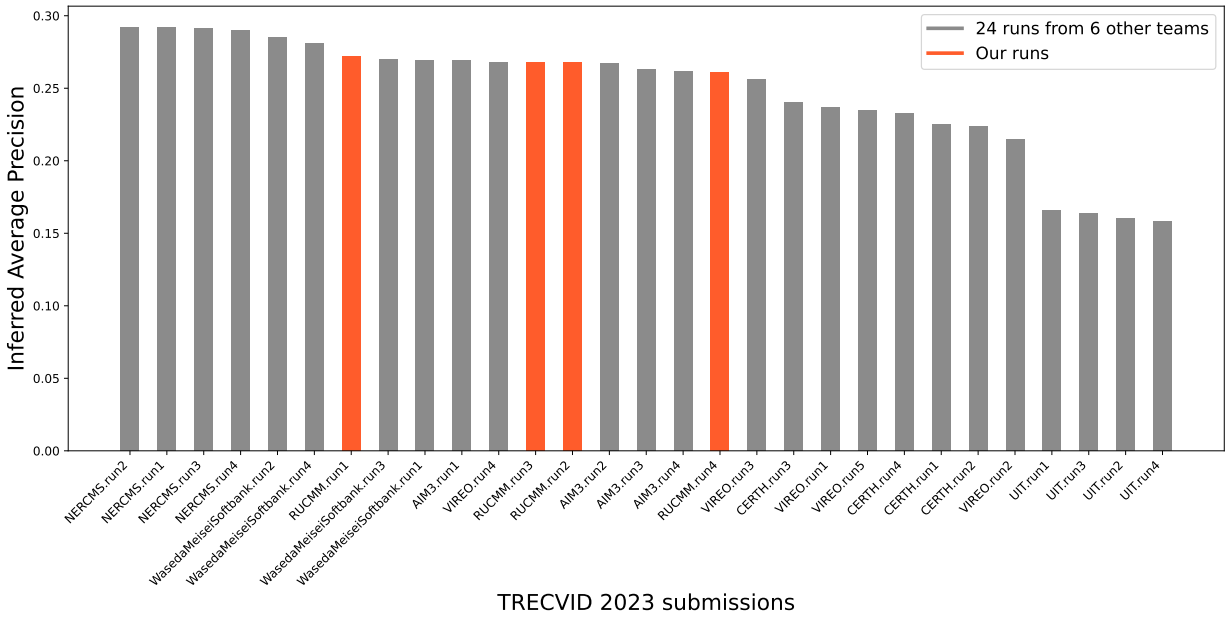

Figure 2: Overview of the TRECVID 2023 AVS benchmark evaluation.

[2] G. Awad, J. Fiscus, D. Joy, M. Michel, A. Smeaton, W. Kraaij, G. Quénot, M. Eskevich, R. Aly, R. Ordelman, G. Jones, B. Huet, and M. Larson. TRECVID 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *TRECVID Workshop*, 2016.

[3] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

[4] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: BERT pre-training of image transformers. In *ICLR*, 2022.

[5] A. Chen, Z. Wang, C. Dong, K. Tian, R. Zhao, X. Liang, Z. Kang, and X. Li. ChinaOpen: A dataset for open-world multimodal learning. In *ACMMM*, 2023.

[6] A. Chen, F. Zhou, Z. Wang, and X. Li. CLIPRerank: An extremely simple method for improving ad-hoc video search. In *ICASSP*, 2024.

[7] D. Ghadiyaram, D. Tran, and M. Feiszli. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, 2019.

[8] F. Hu, A. Chen, Z. Wang, F. Zhou, J. Dong, and X. Li. Lightweight attentional feature fusion: A new baseline for text-to-video retrieval. In *ECCV*, 2022.

[9] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[10] J. Li, D. Li, C. Xiong, and S. C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[11] X. Li, A. Chen, Z. Wang, F. Hu, K. Tian, X. Chen, and C. Dong. Renmin University of China at TRECVID 2022: Improving video search by feature fusion and negation understanding. In *TRECVID Workshop*, 2022.

[12] X. Li, J. Dong, C. Xu, J. Cao, X. Wang, and G. Yang. Renmin University of China and Zhejiang Gongshang University at TRECVID 2018: Deep cross-modal embeddings for video-text retrieval. In *TRECVID Workshop*, 2018.

[13] X. Li, C. Xu, G. Yang, Z. Chen, and J. Dong. W2VV++: Fully deep learning for ad-hoc video search. In *ACMMM*, 2019.

[14] X. Li, J. Ye, C. Xu, S. Yun, L. Zhang, X. Wang, R. Qian, and J. Dong. Renmin University of China and Zhejiang Gongshang University at TRECVID 2019: Learn to search and describe videos. In *TRECVID Workshop*, 2019.

[15] X. Li, F. Zhou, and A. Chen. Renmin University of China at TRECVID 2020: Sentence encoder assembly for ad-hoc video search. In *TRECVID Workshop*, 2020.

[16] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. TGIF: A new dataset and benchmark on animated GIF description. In *CVPR*, 2016.

[17] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[18] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACMMM*, 2022.

[19] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, Y. Bharambe, and L. Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[21] K. Tian, R. Zhao, H. Hu, R. Xie, F. Lian, Z. Kang, and X. Li. TeachCLIP: Multi-grained teaching for efficient text-to-video retrieval. *arXiv*, 2023.

[22] X. Wang, J. Wu, J. Chen, L. Li, Y. F. Wang, and W. Y. Wang. VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019.

[23] J. Xu, T. Mei, T. Yao, and Y. Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[24] H. Zhang, X. Li, and L. Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv*, 2023.