VIREO @ TRECVid 2022 Ad-hoc Video Search

Jiaxin Wu[†], Zhixin Ma[⋆],Chong-Wah Ngo[⋆]

[†]*Department of Computer Science, City University of Hong Kong*
[⋆]*School of Computing and Information Systems, Singapore Management University*
jiaxin.wu@my.cityu.edu.hk,
zxma.2020@phdcs.smu.edu.sg, cwngo@smu.edu.sg

Abstract

In this paper, we summarize our submitted runs and results for Ad-hoc Video Search (AVS) task at TRECVid 2022 [1].

Ad-hoc Video Search (AVS): This year, we applied the hybrid search models ITV [2] and dual-task [3], which enable both concept-free search and concept-based search to retrieve videos. Besides, three new features are integrated into them: 1) a new text-video dataset (i.e., WebVid25M [4]) is appended with previously used video-caption datasets [5, 6, 7] for training data augment; 2) the retrieved result of a text-video pre-training model (i.e., CLIP4Clip [8]) is fused with ITV and dual-task models; 3) a video question answering (video-QA) model [9] is used for video re-ranking. Using different combinations of them, we submitted four automatic runs and four manual runs for the main task. Besides, we also submitted concept runs and novelty run. We briefly summarize our runs as follows:

- *F_D_C_D_VIREO.22_1*: This automatic run attains the mean xinfAP= 0.117 on the main task. It is the ensemble result of the concept-free searches of ITV, dual-task, and the CLIP4Clip models.
- *F_D_C_D_VIREO.22_2*: This automatic run attains the mean xinfAP= 0.110 on the main task. It ensembles the hybrid searches of the ITV and dual-task models.
- *F_D_C_D_VIREO.22_3*: This automatic run obtains the mean xinfAP= 0.142 on the main task. It fuses the retrieved results of the hybrid searches in the ITV and dual-task models and the result of CLIP4Clip. This run prohibits the highest score of all automatic runs.
- *F_D_C_D_VIREO.22_4*: This automatic run attains the mean xinfAP= 0.136 on the main task. It is the re-ranking result of the run *F_D_C_D_VIREO.22_3* using the video-QA model.
- *F_D_C_D_VIREO.22_5*: This concept run attains the mean xinfAP= 0.108 on the main task. It ensembles the concept-based searches of the ITV and the dual-task models.
- *F_D_N_D_VIREO.22_6*: This novelty run is based on the concept-based search of the ITV model with original queries. As a result, this run attains mean xinfAP= 0.088 for the main task.
- *M_D_C_D_VIREO.22_1*: This manual run applies the same system with the same settings presented in the run *F_D_C_D_VIREO.22_1*. The difference is that the original queries are rephrased with new terms. The human intervention improves the performance of the main task from 0.117 to 0.138.

- *M_D_C_D_VIREO.22_2*: This manual run is based on the same system with the same settings presented in the run *F_D_C_D_VIREO.22_2* with manually formulated queries. Similarly, the performance also rises from 0.142 to 0.147 in the main task.
- *M_D_C_D_VIREO.22_3*: This manual run is based on the same setting presented in the run *F_D_C_D_VIREO.22_3* but with manually formulated queries. It improves the automatic result from 0.142 to 0.174.
- *M_D_C_D_VIREO.22_4*: This manual run uses the same system with the same settings presented in the run *F_D_C_D_VIREO.22_4* but with manually formulated queries. The performance rises from 0.136 to 0.166 in the main task.
- *M_D_C_D_VIREO.22_5*: This manual run uses the same system with the same settings presented in the run *F_D_C_D_VIREO.22_5* but with manually formulated queries. However, the performance drops from 0.108 to 0.105 in the main task.

1 Ad-hoc Video Search (AVS)

Our previous effort on interpreting embedding by semantic concepts has obtained great performance on AVS [2, 10]. However, there are still some issues needed to be improved. First of all, our models fall miserably on some queries because of few training instances. To solve it, a large-scale dataset (i.e., WebVid25M [4]) with around 2.5 million text-video pairs is added to the training set. Besides, as recent pre-training models trained on millions of data (e.g., CLIP4Clip [8]) have shown great effectiveness on video retrieval, we also enhance our retrieved video list with this recent technique. Secondly, the existing AVS methods usually wrongly rank near-miss videos on the top list. To overcome this issue, a video-QA model just-ask [9] is used to ask videos some visual questions that are related to the queries. Depending on the answers, the retrieved videos will be re-ranked and the near-miss results are encouraged to be pull down.

1.1 Three enhancements

Besides our ITV model [2] and dual-task model with additional phrase concepts [10], three enhancements are added to solve the above-mentioned data problem and near-miss problem.

1.1.1 WebVid25M dataset

The WebVid25M [4] consists of 2.5 million videos scraped from the Internet and each video is annotated with one caption. The average length of the video is 18 seconds and the videos are open-domain. As the feature extraction is time-consuming, we only used one-fourth of the WebVid25M to append with previously used video caption datasets [5, 6, 7] on training the ITV and the dual-task models. Totally, there are 932,249 video-caption pairs used for training.

1.1.2 CLIP4Clip model

CLIP4Clip [8] is a video-language retrieval model which exploits the knowledge of the pre-trained image-language retrieval model CLIP [11]. Specifically, on the textual side, they directly borrow the text encoder from the CLIP model for query representation. On the visual side, they first extract the CLIP features independently for the sampled frames. Then, CLIP4Clip aggregates the frame features to

Table 1: Samples of generated questions based on tv22 queries

Input query	Output question	Output answer
702 Find shots of blue wall indoors	Where is the blue wall located ?	indoors
	What is the color of the wall ?	blue
711 Find shots of a woman wearing a head kerchief	What type of clothing does a woman wear ?	head kerchief
	Who is the wearing a head kerchief ?	woman
728 Find shots of two adults are seated in a flying paraglider in the air	How many adults are seated in a flying paraglider ?	two
	What type of aircraft are two adults seated in ?	fly paraglider
	Who are seated in a flying paraglider in the air ?	adult

produce the video clip feature and calculates the video-language similarity. They design three approaches for frame feature aggregation and similarity calculation. In this paper, we investigate the performance of the parameter-free approach, which uses mean pooling to aggregate the frame features. To further boost its performance, similar to the ITV [2] and dual-task [3] models, we fine-tune the CLIP4Clip model on MSR-VTT [6], TGIF [5] and VATEX [7] datasets.

1.1.3 Visual question answering

Near-miss videos usually rank high in the retrieval when AVS models cannot differ the detailed differences in videos. Our ITV model has made effort on this issue by avoiding the co-existence of contrary videos, such as indoor/outdoor clips, using unlikelihood learning. However, it only solved part of the scope. Here, a video question answering (video-QA) model is used to prune similar but incorrect videos in a more general way. Below, we describe how to use video-QA for re-ranking.

First of all, we use two question generation models [12, 13] to produce question-answer pairs based on the AVS query. Specifically, [12] is a rule-based model which designs several rules to generate questions based on the syntactic structure of the input sentence. [13] is an end-to-end transformer model which learns question generation from a large annotated question-answer dataset. As a result, 199 questions are generated for 30 tv22 queries. Table 1 shows some samples of the generated question-answer (QA) pairs. For example, for the input AVS query-702 *Find shots of blue wall indoors*, two QA pairs are produced by asking about location and color.

After having QA pairs, a state-of-the-art video-QA model [9] is used to get predicted answers of video by using the generated questions. Specifically, just-ask [9] is an end-to-end transformer model pre-trained on a large video-QA dataset. If the predicted answer matches the expected answer, the video will be re-ranked in a higher position than those mismatched videos. In tv22, we re-rank the top 200 videos.

2 Results analysis

In this year's AVS benchmarking, the evaluation is conducted on the V3C2 dataset [14] with 30 queries.

2.1 Impact of the WebVid25M dataset

Table 2 compares the performances of training models with and without WebVid25M dataset. The result shows that adding this large dataset significantly improves the xinfAP scores on all search modes in both two models. Specifically, more than half of tv22 queries have better performances on every search mode. For example, for the query-709 *Find shots of a person is in the act of swinging*, more correct video clips are found on both concept-based and concept-free searches in the ITV model because of more

Table 2: Performance comparison of models trained with and without WebVid25M dataset on tv22 (mean xinfAP)

Model	search mode	w/o WebVid25M	w/ WebVid25M
ITV	concept-based search	0.077	0.102
	concept-free search	0.087	0.119
	hybrid search	0.103	0.131
dual-task with additional phrase concepts	concept-based search	0.050	0.077
	concept-free search	0.073	0.126
	hybrid search	0.093	0.137

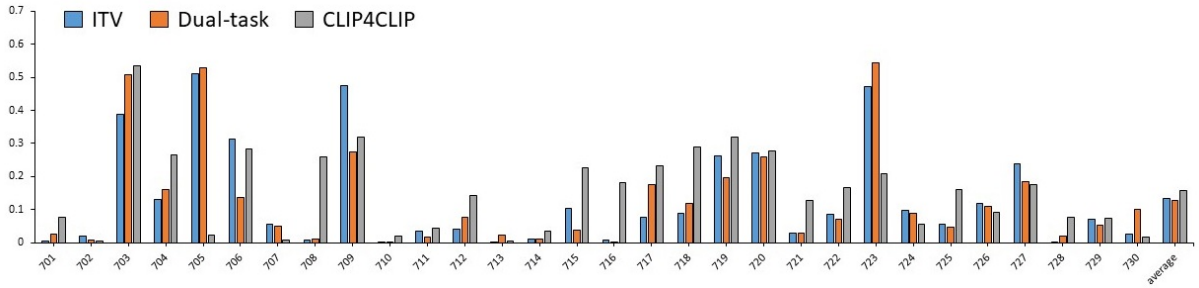

Figure 1: Performance comparison of CLIP4Clip and other components of our run3 on tv22 query set.

training instances on swinging. In fact, their performances are three and four times higher than the originals, and the xinfAP score of hybrid search is improved from 0.164 to 0.475.

2.2 Impact of CLIP4Clip model

Comparing the performances of our run2 and run3, the integration of the CLIP4Clip model brings a 29.1% improvement in automatic queries. Specifically, run3 outperforms run2 on 26 out of 30 queries. Figure 1 further shows the performances of each component in the run3. CLIP4Clip obtains the highest scores on 20 out of 30 queries. Especially on those hard queries such as query-708 *Find shots of a female person bending downwards*, query-710 *Find shots of a person wearing a light t-shirt with dark or black writing on it*, and query-716 *Find shots of a drone landing or rising from the ground*, CLIP4Clip finds many correct video clips while ITV and dual-task barely find some. The pre-trained model CLIP4Clip significantly outperforms ITV and dual-task models on the tv22 query set by taking advantage of pre-training on huge data.

2.3 Impact of the video-QA model

We first analyze whether the video-QA model makes a change to the rank list. The average position change of a video is computed as follows:

$$c = \frac{1}{N}\sum_{i}^{N} re\_rank(v_i) - ori\_rank(v_i),$$

where N is the number of video for re-ranking. Here, N is equal to 200 in this paper. The result shows the average change c per query is about 50, which indicates the video-QA model brings much change to the list.

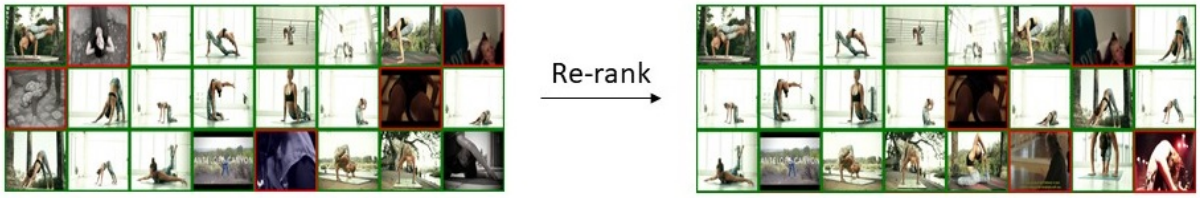

Figure 2: Re-ranking result of the query-708 *Find shots of a female person bending downwards.* xinfAP rises from 0.231 to 0.298.

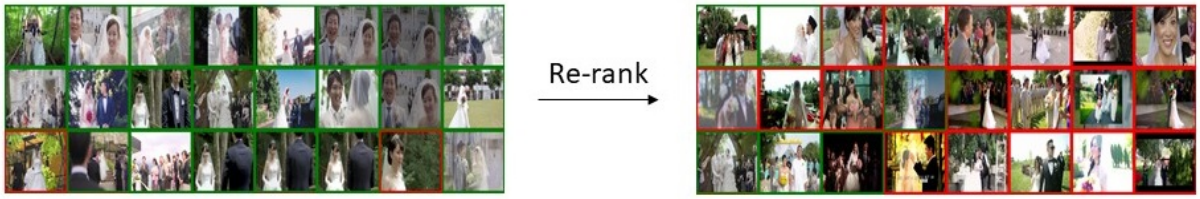

Figure 3: Re-ranking result of the query-715 *Find shots of an Asian bride and groom celebrating outdoors.* xinfAP drops from 0.373 to 0.220.

We also compare the xinfAP score with a search length of 200 between run3 and run4 to investigate the impact on performance. With the involvement of the video-QA model, the mean xinfAP@200 drops from 0.244 to 0.226. Specifically, the re-ranking brings improvement to 12 out of 30 queries. However, 17 queries have worse scores and one query is unchanged. Figure 2 visualizes the original and updated rank lists of an improved query *Find shots of a female person bending downwards.* The related question-answer pair is "what gender is the person bending downwards"-"female". The performance of this query is elevated as videos of man are pushed down. In contrast, Figure 3 shows a drop case where the video-QA model asks "where would a bride and groom celebrate?" and the model could not properly answer "outdoors" for those correct video clips on the original top list.

2.4 Impact of the manually modified queries

This year, our manual runs are able to outperform our automatic runs. Table 3 lists the original and modified queries and their performances. 22 out of 30 queries have a performance boost, especially those hard queries. For example, after modifying the query-710 *A person wearing a light t-shirt with dark or black writing on it* to *A person wears a white t-shirt with black letters*, the xinfAP is elevated from 0.008 to 0.116. However, for those negation queries that have the prefix non- to retrieve the opposite, the modified queries did not work, such as query-721 and query-730. We attribute this to our manual queries being somewhat trying luck or arbitrary.

3 Conclusion

Our study this year aims to address the limitations of training data and near-miss videos in our ITV and dual-task models. First, the appended WebVid25M dataset is verified to be effective in improving performance by bringing more training instances to the ITV and the dual-task models. Second, the pre-trained model CLIP4Clip shows superior performance on most of the tv22 queries and improves its ensemble of the ITV and dual-task models. Finally, using video-QA brought a large change to the rank

Table 3: Performance comparison of the original queries and the modified queries, i.e., *F_D_C_D_VIREO.22_3* versus *M_D_C_D_VIREO.22_3*.

original query	xinfAP	modified query	xinfAP
701 A man with a white beard	0.048	701 Old man with white beard	0.303
702 A room with blue wall	0.012	702 Blue wall indoors	0.019
703 A construction site	0.353	703 Tower crane, workers on construction site	0.346
704 A parked white car	0.243	704 White car in parking lot	0.305
705 A type of cloth hanging on a rack, hanger, or line	0.444	705 Hanging cloth	0.157
706 Building with columns during daytime	0.275	706 Building with columns outdoors during daytime	0.304
707 A person is mixing ingredients in a bowl, cup, or similar type of containers	0.045	707 Hand mixing food in a bowl	0.088
708 A female person bending downwards	0.097	708 A woman practicing yoga bending downwards	0.294
709 A person is in the act of swinging	0.258	709 A person playing trapeze or swing	0.441
710 A person wearing a light t-shirt with dark or black writing on it	0.008	710 A person wears a white t-shirt with black letters	0.116
711 A woman wearing a head kerchief	0.015	711 A lady wears a head bandana	0.051
712 A man wearing black shorts	0.118	712 A man wears black shorts outdoors	0.122
713 A kneeling man outdoors	0.013	713 Man knee on ground	0.006
714 Two or more persons in a room with a fireplace	0.009	714 Family talking with fireplace in the living room	0.051
715 An Asian bride and groom celebrating outdoors	0.123	715 Asian bride and groom celebrating their marriage outdoors	0.13
716 A drone landing or rising from the ground	0.001	716 A drone on the ground	0.097
717 A black bird seen on a dry area sitting, walking, or eating	0.149	717 A black bird on the ground	0.192
718 A large stone building from the outside	0.292	718 A stone building outdoors	0.182
719 A piece of heavy farm equipment or machine seen outdoors	0.18	719 Tractors on farmyard	0.233
720 A clock on a wall in a room	0.217	720 Hanging clock on wall indoors	0.259
721 Two persons are seen while at least one of them is speaking in a non-English language outdoors	0.094	721 Two asian people talking outdoors	0.055
722 A woman is eating something outdoors	0.076	722 A woman eating or biting food outdoors	0.116
723 A person is biking through a path in a forest	0.449	723 A person riding a bicycle on a forest trail	0.699
724 A man and a bike in the air after jumping from a ramp	0.171	724 A man jumping with bicycle from a ramp	0.057
725 A woman holding or smoking a cigarette	0.053	725 A female with a cigarette	0.088
726 Two teams playing a game where one team have their players wearing white t-shirts.	0.193	726 Two teams are playing a sport game with white cloth	0.197
727 Two persons wearing white outfits and black belts demonstrate martial arts in a room with floor mats	0.114	727 Two persons doing judo demonstration	0.147
728 Two adults are seated in a flying paraglider in the air	0.017	728 Two persons on a paraglider seat with feet visible	0.024
729 A ring shown on the left hand of a person	0.13	729 A ring on the left hand	0.111
730 A man is holding a knife in a non-kitchen location	0.077	730 A man holding a knife outdoors	0.034
average	0.142		0.174

list. However, the change did not boost the performance on every query, as the video-QA model is not powerful enough to answer every question correctly.

For the manual run, the problem where the results are sensitive to how a query is expressed remains. Our result of the manual run is somewhat "trying luck" or arbitrary.

4 Acknowledgments

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

References

[1] G. Awad, K. Curtis, A. A. Butt, J. Fiscus, A. Godil, Y. Lee, A. Delgado, J. Zhang, E. Godard, B. Chocot, L. Diduch, J. Liu, Y. Graham, , and G. Quénot, "An overview on the evaluated video retrieval tasks at trecvid 2022," in *Proceedings of TRECVID 2022*. NIST, USA, 2022.

[2] J. Wu, C.-W. Ngo, W.-K. Chan, and Z. Hou, "(Un)likelihood training for interpretable embedding," 2022.

[3] J. Wu and C.-W. Ngo, "Interpretable embedding for ad-hoc video search," in *Proceedings of the ACM Conference on Multimedia*, 2020.

[4] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *IEEE International Conference on Computer Vision*, 2021, pp. 1–15.

[5] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[6] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[7] 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 *The IEEE International Conference on Computer Vision*, 2019.

[8] 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," *arXiv preprint arXiv:2104.08860*, pp. 1–14, 2021.

[9] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1686–1697.

[10] J. Wu, Z. Hou, Z. Ma, and C.-W. Ngo, "VIREO@trecvid 2021: Ad-hoc video search," in *In NIST TRECVID Workshop*, 2021.

[11] 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.

[12] M. Ren, R. Kiros, and R. S. Zemel, "Exploring models and data for image question answering," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015, p. 2953–2961.

[13] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485—-5551, 2020.

[14] L. Rossetto, H. Schuldt, G. Awad, and A. A. Butt, "V3C–a research video collection," in *International Conference on Multimedia Modeling*. Springer, 2019, pp. 349–360.