Text Query to Web Image to Video: A Comprehensive Ad-hoc Video Search

Supplementary Material

Nhat-Minh Nguyen[1,2], Tien-Dung Mai[1,2], and Duy-Dinh Le[1,2]

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
21521135@gm.uit.edu.vn
{dungmt,duyld}@uit.edu.vn

In this supplementary material, we provide more results, visualizations, and in-depth discussions about Image Search Engine Results and Retrieval Results.

1 Image Search Engine Results

Consider the top images returned by Bing Images for query ID 732 in the TRECVID2023 AVS main task query set, with the content *"A woman with red hair"*, shown in Fig. 1. We can observe that all result images displayed in Fig. 1 satisfy the query description requirements: "a woman" and "with red hair". Similarly, for query ID 709 in TRECVID2022 AVS main task query set, "A person is in the act of swinging", shown in Fig. 2, the majority of results returned by Yahoo Images satisfy the query description: "a person" and "is in the act of swinging". Except for the last image, there are two people are in the act of swinging. Images returned by the Image Search Engine include both real-life and non-real-life images that satisfy the query.

Consider the top results returned by Bing Images for a different query, ID 748 in the TRECVID2023 AVS main task query set, with the content *"A man carrying a bag on one of his shoulders (excluding backbags)"*, shown in Fig. 3. In this case, irrelevant results are more prevalent, comprising the majority of the results displayed in Fig. 3. Most of these images depict a man carrying a bag on one shoulder, however, the majority are backpacks. This means that most results only satisfy 2 out of 3 query descriptions: "a man" and "carrying a bag on one of his shoulders", without satisfying the requirement "excluding backpacks". This may be due to the search engine not fully comprehending the entire query request, as well as the semantics in the displayed images, leading to the return of images that do not fully satisfy the query.

This demonstrates that using complex/detailed descriptions of the desired image in queries can lead to image search engines not fully grasping the query's content, resulting in confusion during the process of identifying images that satisfy the requirements of the image search engines. Other queries with complex descriptions, leading to many irrelevant images being returned by image search engines including: *"A man is talking in a small window located in the lower corner of the screen"*, *"Two persons are seen while at least one of them is speaking*

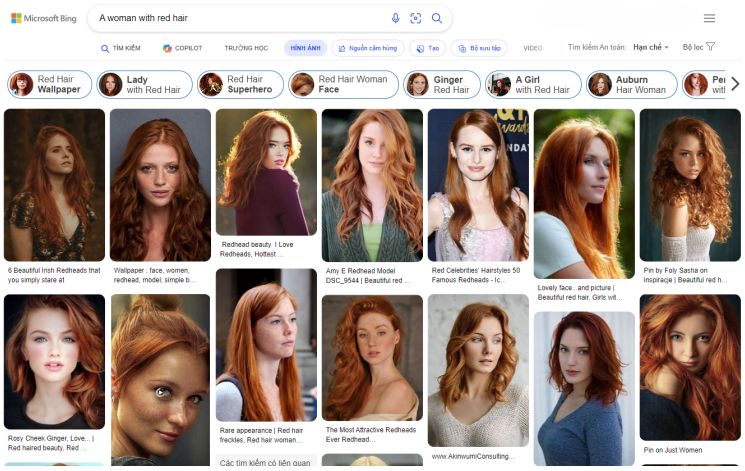

Fig. 1: The top result images returned by Bing Images for query ID 732 in the TRECVID2023 AVS main task query set, with the content: *"A woman with red hair"*.

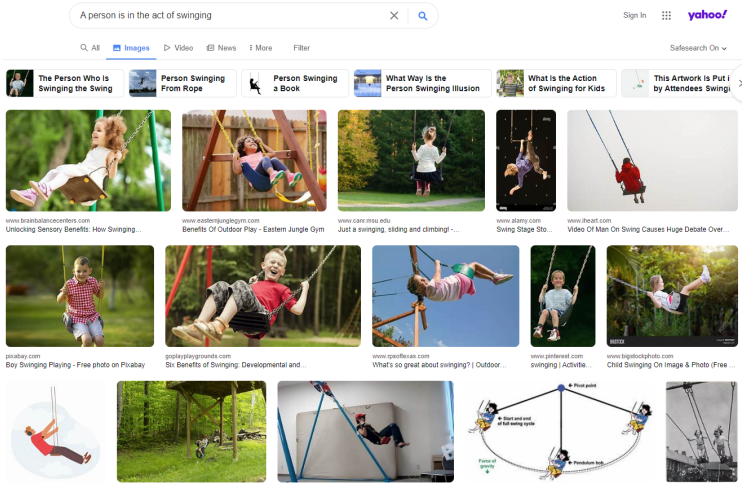

Fig. 2: The top result images returned by Yahoo Images for query ID 748 in the TRECVID2022 AVS main task query set, with the content: *"A person is in the act of swinging"*.

in a non-English language outdoors", *"Two teams playing a game where one team have their players wearing white t-shirts"*,...

2 Retrieval Results

The top video segments returned by our proposed method for query ID 703 in TRECVID2022 AVS main task query set, with the content *"A construction*

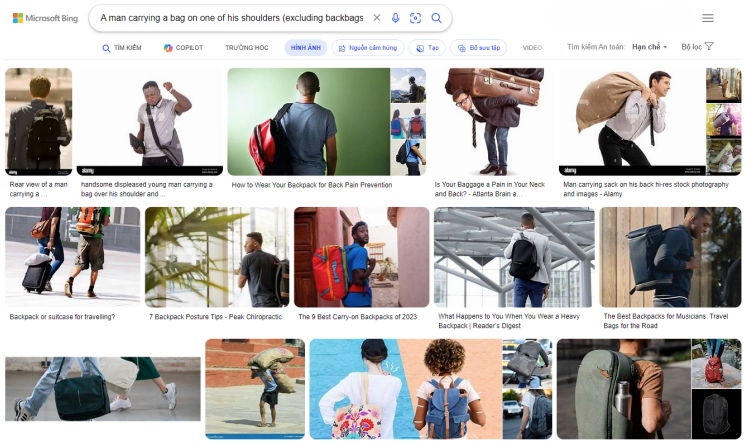

Fig. 3: The top result images returned by Bing Images for query ID 748 in the TRECVID2023 AVS main task query set, with the content: *"A man carrying a bag on one of his shoulders (excluding backbags)"*.

site", are fully consistent with the ground truth (all have green borders). The video segments, represented by their keyframes (the images retrieved using the query), are shown in Fig. 4. The performance metrics for the results returned for this query are xinfAP = 0.7359, AP = 0.8302, as shown in Table 1. It means the results returned by our method for this query are highly accurate (considering the top results).

Consider the top result video segments returned by our method for query ID 709 in TRECVID2022 AVS main task query set, with the content: *"A person is in the act of swinging"*, shown in Fig. 5. The vast majority of returned results are satisfactory and consistent with the ground truth. Examining the 6 incorrect results (with red borders), our method exhibits confusion when these keyframes have limited viewing angles, only capturing a part of the object. However, this confusion accounts for a small portion of the results displayed in Fig. 5. Similarly, for query ID 732 in the TRECVID2023 AVS main task query set, with the content: *"A woman with red hair"*, the vast majority of the top results returned satisfy the query and are consistent with the ground truth, shown in Fig. 6. However, there are 5 incorrect results in Fig. 6, which are due to the feature extraction model's confusion in identifying hair color (influenced by objective factors: surrounding environment, image color) and blurred frames where identification is not possible. Nevertheless, these incorrect results are few, and the top results returned by the proposed system for this query and the query "A person is in the act of swinging" are still sufficiently good.

Besides the queries for which our system returns good results, there are still queries with poor results. Consider query ID 748 in TRECVID2022 AVS main task query set, with the content: *"A man carrying a bag on one of his shoulders (excluding backbags)"*. The keyframes of the top video segments returned by the

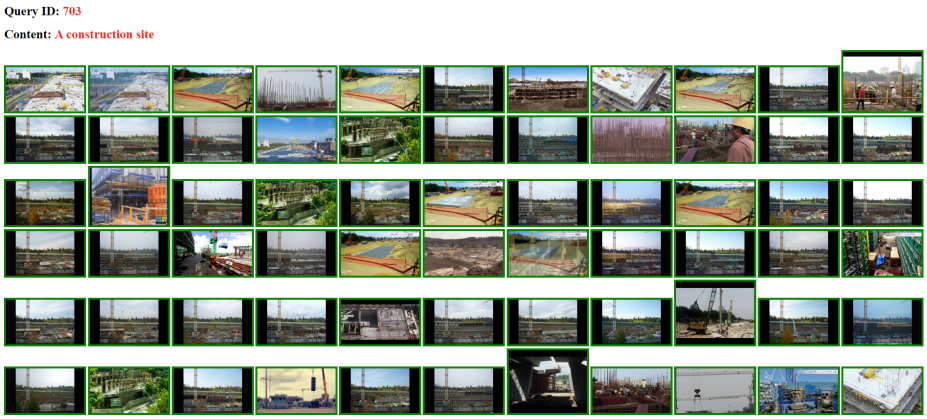

Fig. 4: Visualize the top query results with keyframes of each video segment returned for query ID 703 in the TRECVID2022 AVS query set when combining CLIP and BEiT-3 for feature extraction, integrating query images from Yahoo Images and text query. Query content is *"A construction site"*. Keyframes with green borders represent video segments that match the ground truth, while those with red borders do not.

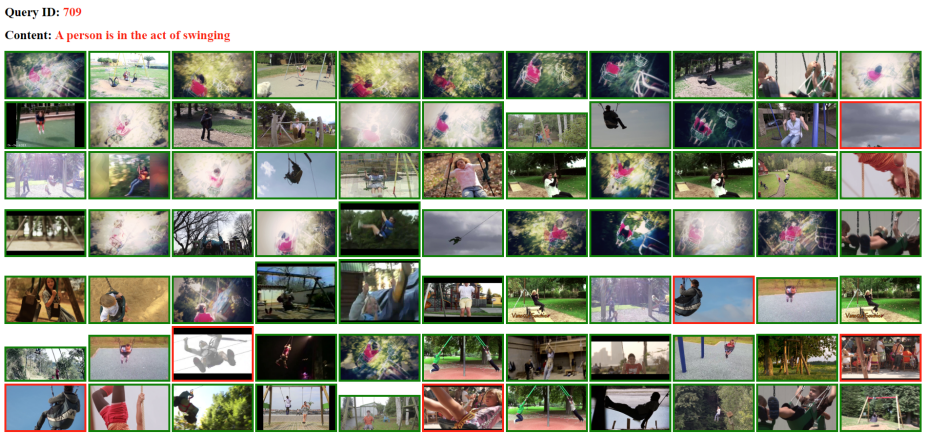

Fig. 5: Visualize the top query results with keyframes of each video segment returned for query ID 709 in the TRECVID2022 AVS query set when combining CLIP and BEiT-3 for feature extraction, integrating query images from Yahoo Images and text query. Query content is *"A person is in the act of swinging"*. Keyframes with green borders represent video segments that match the ground truth, while those with red borders do not.

system, shown in Fig. 7, are mostly incorrect compared to the ground truth. The displayed keyframes mostly only satisfy 2 descriptions in the query, "A man" and "carrying a bag on one of his shoulders," but not the description "excluding backpacks." This is due to two factors: the query images obtained

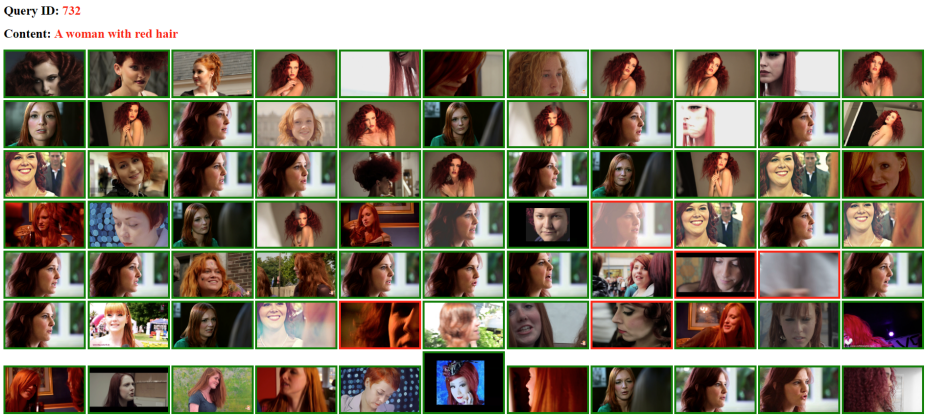

Fig. 6: Visualize the top query results with keyframes of each video segment returned for query ID 732 in the TRECVID2023 AVS query set when combining CLIP and BEiT-3 for feature extraction, integrating query images from Bing Images and text query. Query content is *"A woman with red hair"*. Keyframes with green borders represent video segments that match the ground truth, while those with red borders do not.

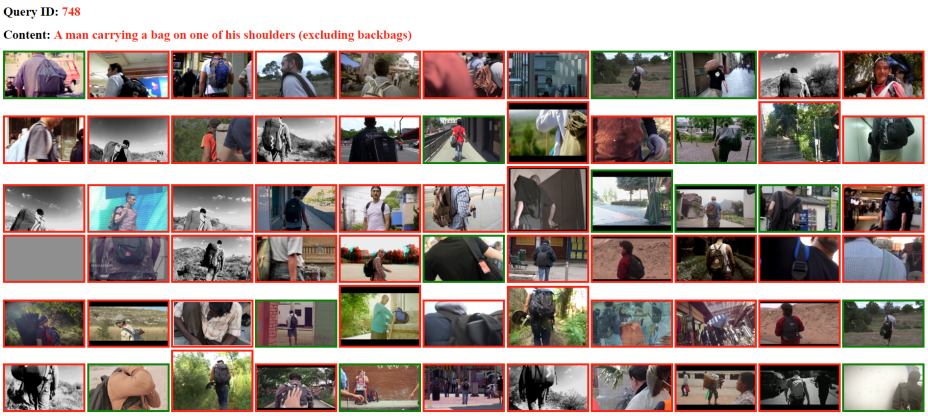

Fig. 7: Visualize the top query results with keyframes of each video segment returned for query ID 748 in the TRECVID2023 AVS query set when combining CLIP and BEiT-3 for feature extraction, integrating query images from Bing Images and text query. Query content is *"A man carrying a bag on one of his shoulders (excluding backbags)"*. Keyframes with green borders represent video segments that match the ground truth, while those with red borders do not.

from the image search engine also return similar incorrect results, as shown in Fig. 3, and the feature extraction model does not fully grasp the entire query description.

Observe the graphs comparing the system's evaluation results when using images from Google Images, Bing Images, and Yahoo Images with the mean

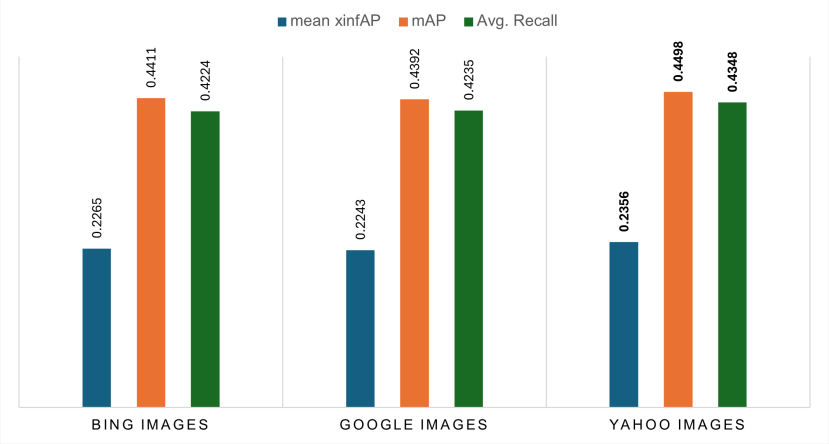

Fig. 8: Comparison of query results using query images obtained from 3 different image search engines (combining CLIP and BEiT-3 for feature extraction, integrating query images and text query) for the TRECVID2022 AVS main task query set.

Fig. 9: Comparison of query results using query images obtained from 3 different image search engines (combining CLIP and BEiT-3 for feature extraction, integrating query images and text query) for the TRECVID2023 AVS main task query set.

xinfAP, mAP, and Avg. Recall metrics, shown in Fig. 8 for the 2022 query set and Fig. 9 for the 2023 query set. Observing the graph for the 2022 query set shown in Fig. 8, the proposed system, when utilizing query images aggregated from Yahoo Images, yields the highest results, with a considerable difference (but less than 0.01) compared to Google Images and Bing Images in terms of mean xinfAP and mAP metrics. For the graph of the 2023 query set shown in Fig. 9, the evaluation results for all three metrics do not exhibit significant differences

when using query images from the 3 image search engines for the retrieval system. This demonstrates that the system performs relatively consistently when using different image search engines, meaning there is no significant difference.

Table 1: Detailed evaluation for queries in the TRECVID2022 AVS main task query set. The results used for evaluation are the results returned when combining CLIP and BEiT-3 for feature extraction, integrating query images from Yahoo Images and text query for retrieval.

	TRECVID2022 AVS		
Query ID	xinfAP	Avg. Precision	Recall
701	0.2214	0.5525	0.5170
702	0.0625	0.4326	0.2000
703	0.7359	0.8302	0.5591
704	0.2731	0.5117	0.3162
705	0.3666	0.7232	0.4478
706	0.3392	0.6503	0.3642
707	0.0767	0.2289	0.4829
708	0.2333	0.5167	0.2963
709	0.4942	0.7518	0.3063
710	0.1040	0.3264	0.3882
711	0.0700	0.1589	0.4366
712	0.0381	0.2668	0.1179
713	0.0581	0.2698	0.2929
714	0.0878	0.2489	0.4479
715	0.2452	0.5604	0.4711
716	0.2702	0.3578	0.8165
717	0.1443	0.3318	0.5125
718	0.3855	0.6022	0.2795
719	0.3372	0.5581	0.7037
720	0.2928	0.6457	0.5780
721	0.0458	0.2059	0.1430
722	0.1587	0.3174	0.5504
723	0.6886	0.7755	0.3137
724	0.3302	0.6123	0.5633
725	0.1787	0.4778	0.4914
726	0.0396	0.1401	0.0844
727	0.2731	0.2750	0.9863
728	0.2500	0.4483	0.6339
729	0.2311	0.5701	0.4968
730	0.0346	0.1462	0.2447
mean	**0.2356**	**0.4498**	**0.4348**

Deeper analysis of the evaluation results for each query in the 2022 and 2023 query sets, shown in Table 1 and Table 2, respectively. With detail results of 2022 query set, shown in Table 1, the top 5 queries with the highest xinfAP

Table 2: Detailed evaluation for queries in the TRECVID2023 AVS main task query set. The results used for evaluation are the results returned when combining CLIP and BEiT-3 for feature extraction, integrating query images from Bing Images and text query for retrieval.

	TRECVID2023 AVS		
Query ID	xinfAP	Avg. Precision	Recall
731	0.4625	0.8313	0.4782
732	0.4026	0.7256	0.4746
733	0.3456	0.6161	0.6477
734	0.3522	0.5648	0.3689
735	0.2466	0.6722	0.4481
736	0.0664	0.1985	0.3620
737	0.6326	0.7954	0.3866
738	0.1589	0.3722	0.5203
739	0.0535	0.1819	0.3976
740	0.3405	0.5573	0.2389
741	0.0226	0.1719	0.1916
742	0.1217	0.2346	0.5909
743	0.0003	0.0151	0.0435
744	0.1291	0.3457	0.4348
745	0.4495	0.7257	0.2039
746	0.2299	0.4917	0.5944
747	0.1419	0.4333	0.4382
748	0.0352	0.1755	0.2506
749	0.7277	0.8255	0.3521
750	0.0429	0.2220	0.2528
mean	**0.2481**	**0.4578**	**0.3838**

scores are ID 703, 723, 709, 718, 705 (xinfAP > 0.36), and the top 5 queries with the lowest xinfAP scores are ID 730, 712, 726, 721, 713 (xinfAP < 0.06), with 9/30 queries having xinfAP < 0.1. Although many queries have low xinfAP scores, with the Avg. Precision metric, all queries are higher than 0.14, with the lowest value for query ID 726 (AP = 0.1401) and the highest for query ID 723 (AP = 0.7755), with mean AP = 0.4498. The lowest Recall is 0.0844 for query ID 726 and the highest is 0.9863 for query ID 727, with mean of Recall (Avg. Recall) is 0.4348.

With the detailed evaluation results of the 2023 query set, shown in Table 2, the top 5 queries with the highest xinfAP scores are ID 749, 737, 731, 745, 732 (xinfAP > 0.40), and the top 5 queries with the lowest xinfAP scores are ID 743, 741, 748, 750, 739 (xinfAP < 0.06), with 6/20 queries having xinfAP < 0.1. Except for query ID 743, which has a value lower than 0.05 on all three metrics, the remaining queries have an AP higher than 0.17 and Recall higher than 0.19. The highest value achieved is for query ID 731 with AP = 0.8313 on the AP metric, and the highest Recall value achieved is for query ID 733 with Recall = 0.6477.

In Table 1, 2, some queries have Recall values much higher than xinfAP and mAP, notably query ID 727 in the 2022 query set. The low AP is due to relevant video segments not being ranked high in the returned results list, while the high Recall is because the returned results list contains almost all relevant video segments in the ground truth. Furthermore, the number of relevant video segments in the ground truth for this query ID is only 73 [1], while the returned results list contains up to 1000 video segments, leading to xinfAP and AP values much lower than Recall. Conversely, some queries have xinfAP and AP values much higher than Recall, notably query ID 745 in the 2023 query set. This is because this query has many video segments in the dataset that are relevant to the ground truth, while the number of results returned by the system is limited (1000 video segments). Specifically, the number of video segments considered relevant to query ID 745 among the evaluated video segments of the V3C2 dataset is 2496 [2], approximately 2.5 times higher than the maximum number of results returned by the system (1000).

Considering query ID 743 in the 2023 query set, with the content "A man is talking in a small window located in the lower corner of the screen," this is the most challenging query in the 2023 query set. The results for all research groups in the TRECVID 2023 AVS task have a median xinfAP < 0.001 [2]. Our system's evaluation for this query also yields very low results with xinfAP $= 0.0003$, AP $= 0.0151$, and Recall $= 0.435$.

References

1. Awad, G., Curtis, K., Butt, A., Fiscus, J., Godil, A., Lee, Y., Delgado, A., Godard, E., Diduch, L., Liu, J., et al.: An overview on the evaluated video retrieval tasks at trecvid 2022. arXiv preprint arXiv:2306.13118 (2023)
2. Awad, G., Curtis, K., Butt, A.A., Fiscus, J., Godil, A., Lee, Y., Delgado, A., Godard, E., Diduch, L., Gupta, D., et al.: Trecvid 2023–a series of evaluation tracks in video understanding. In: Proceedings of TRECVID. vol. 2023 (2023)