Leveraging LLMs and Generative Models for Interactive Known-Item Video Search

Zhixin MA[1], Jiaxin WU[1,2], and Chong Wah NGO[1]

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

Abstract. While embedding techniques such as CLIP have considerably boosted search performance, user strategies in interactive video search still largely operate on a trial-and-error basis. Users are often required to manually adjust their queries and carefully inspect the search results, which greatly rely on the users' capability and proficiency. Recent advancements in large language models (LLMs) and generative models offer promising avenues for enhancing interactivity in video retrieval and reducing the personal bias in query interpretation, particularly in the known-item search. Specifically, LLMs can expand and diversify the semantics of the queries while avoiding grammar mistakes or the language barrier. In addition, generative models have the ability to imagine or visualize the verbose query as images. We integrate these new LLM capabilities into our existing system and evaluate their effectiveness on V3C1 and V3C2 datasets.

Keywords: Large Language Models · Generative Model · Known-Item Search · Interactive Video Retrieval

1 Introduction

Advancements in cross-modal embedding techniques have substantially enhanced the performance of video searches on large-scale datasets. In the annual competition of interactive video search, e.g., video browser showdown (VBS) [5,2], leveraging a pre-trained transformer-based joint embedding models (e.g., CLIP [9], CLIP4Clip [6] and BLIP [4]) has become a norm in recent years. For example, the top-performance team [12] in VBS2023 has used CLIP [9] features to measure the similarity of the input textual query and videos. Nevertheless, individual user proficiency still has a great impact on search performance. In the current VBS, the users have to inspect the return search results after initializing a search query. When the return results are not satisfying, users need to revise the query and browse the results repetitively. The effect of such trial-and-error search also

depends on personal language proficiency and visual sensitivity, making search results user-dependent as also studied in [7].

In this paper, we propose using LLMs to assist users in query formulation. In recent years, LLMs such as GPT-4 [8] and LLaMA [13] have demonstrated surprising capability in natural language understanding and generation. The involvement of LLMs in interactive systems can help users improve the quality of query formulation, like correcting grammar mistakes, and diversify the query semantics by rephrasing the query in multiple ways. Table 1 shows the queries of TRECVid Ad-hoc Video Search (AVS) and VBS textual Known-Item Search (KIS) which are rephrased by GPT-4. The post-fix "-rx" indicates the x-th rephrased caption. The upper section of Table 1 compares the original and rephrased captions of query 731 from AVS 2023. The first rephrased caption keeps the keywords of the original query (i.e., a man and a baby) and expresses "is seen" in a different way. The second one details the man as a gentleman and the baby as an infant. In the example of the KIS query, the original description includes three shots. However, the current generative models cannot produce successive frames for multiple shots described in a video caption. Therefore, we designed a prompt to summarize the query and expect the LLM can find the key information to index the target query. As shown in the lower section of Table 1, the GPT-4 selects the last shot from the KIS query and rephrases it in two different ways, which helps the search engine focus on a specific scene. In the interactive system, a pop-up window is employed to display the GPT-rephrased queries and allow user to evaluate and pick a query as the search engine's input.

In addition to the query formulation, having a good understanding of the textual query in the KIS task and envisioning the target video in mind are challenging for users, especially those people who are not native English speakers. Inspired by the generative models [10] capability in understating natural language descriptions and generating high-quality images, we propose to use generative models to perceive the information need and do the imagination for users in the KIS task. Specifically, given a detailed caption describing the target video, we input the caption to a generative model (e.g., stable diffusion [10]) to generate images. The generated images are subsequently used in two ways. On one hand, the generated image can be directly employed as a visual query to optimize the rank of target shots by similarity search. On the other hand, the visualized search targets can inspire users to use them as references during the results inspection. In a search session, the user query will be fed into the generative model. The generated images will be displayed in the interface to help the user envision the target scene. Thereafter, the user can pick the generated images as a visual query to optimize the ranked list.

We conduct preliminary experiments on the V3C1 [1] and V3C2 [11] datasets to demonstrate the impact of engaging LLMs and generative models in the interactive search. Specifically, we use the visual query generated by the stable diffusion model [10] to improve the mAP performance on the TRECVid AVS task and examine the effectiveness of the LLM-rephrased query in indexing the search target on the textual KIS task. Furthermore, we enhance the ITV [14]

Table 1: Comparison between the original and rephrased query by GPT-4. The post-fix "-rx" indicates the x-th rephrased caption. The upper section shows the query comparison from the AVS task and the lower section presents a more detailed query from VBS textual KIS task.

Query ID	Description
tv23-731	A man is seen with a baby
tv23-731-r1	A man and a baby are in the shot
tv23-731-r2	A gentleman is spotted with an infant
vbs23-kis-t6	A sequence of three shots: two people and a wall with posters, a balcony with laundry hanging on a rope and a train passing behind two standing tombstones. In the first shot, a person is standing around a corner of the wall on the right, the other person walks away to the left. In the last shot, there is a fence between the graveyard and the railway line, and the roof of a building is visible behind the railway line.
vbs23-kis-t6-r1	A train moves past two tombstones, separated by a fence from the railroad, with a building's roof in the background.
vbs23-kis-t6-r2	Two tombstones stand tall as a train goes by, a fence in the forefront and the silhouette of a building behind.

feature in our system and achieve competitive performance compared to the state-of-the-art transformer-based methods.

2 Textual Query Enhancement

In this section, we elaborate on the engagement of the LLMs and generative models in our interactive video search system. Specifically, for a user query, we employ the LLM (i.e., GPT-4 [8]) to rephrase the textual query and generative model (i.e., stable diffusion [10]) to generate visual queries.

2.1 Textual Query Re-formulation with LLMs

We employ GPT-4 [8] to rephrase the query to improve the user query's quality and robustness. As shown in Table 1, the two rephrased queries, ending with "-r1" and "-r2", present the same semantics in different expressions. The prompt for query generation is "The following lines show a list of video captions. Please rephrase each video caption in 5 different ways. If there are multiple shots or scenes in a caption, please select the most distinguishable one and avoid describing multiple scenes in a caption". In our interactive system, the prompt is preset. Once the user inputs a textual query, the system will pass the query as well as the pre-defined prompt to GPT4 API and display the refined captions in a pop-up window. The users can evaluate the goodness of the candidates and select one caption for search.

2.2 Visual Query Generation with Generative Model

In the textual KIS task, given the textual description of the search target, users' imagination of the target has a great impact on the browsing. It is challenging to browse the search results and identify the target from the search rank list without a picture in mind. However, it is easy for users to select a few images

Fig. 1: Examples of the generated images for two VBS Textual KIS queries (i.e., VBS22-KIS-t3), along with their target videos shown on the left.

imagined by the diffusion model that are close to their visual impression of the search target. The selected images are then used as a query for search. In this paper, we propose to use the generative model (e.g., stable diffusion [10]) to do the imagination for the user, i.e., generating images based on the given textual queries. Given a query, three images are generated with different seeds using the stable diffusion model. Figure 1 shows examples of the image generations for two VBS textual KIS queries, i.e., VBS22-KIS-t3 [3] , along with their search targets. As seen, the target video and the generated images are visually similar. In the system, we employ the generated images in two ways. Firstly, they will be displayed to the user as a visual guide to identify the search target. Moreover, the users could select a subset of the generated images as a visual query to find visually similar videos in our new interactive retrieval system.

3 Experiments

To evaluate the quality of the rephrased textual query and generated visual query in video search, we compare them with using original queries to search. The experiments are conducted on the V3C1 [1] and V3C2 [11] datasets. The text-video similarity is measured by three models, i.e., CLIP4Clip [6], BLIP2 [3], and enhanced ITV [14]. In the evaluation, we report mean average precision (mAP) on the TRECVid AVS tasks and the rank of the search target on VBS textual KIS tasks.

Table 2 presents the rank of the search target using different query types. When multiple shots are included in the target segment, we employ the best shot rank. Models ending in "-t", "-r", and "-u" denote the original, GPT-rephrased, and the user query, respectively. The suffix "-v3" refers to the mean pooling of three generated images. The user queries of "-u" are collected during the search session where we ask a novice user to locate the target in the real KIS setting using the CLIP4Clip feature. The * in the cells indicates the user finally finds the correct target. As depicted in Table 2, almost all the found queries are ranked within the top-100 results. In terms of the target rank, the user performs similarly with GPT in VBS 22 but achieves a higher rank on more

[3] VBS23-KIS-t3: Almost static shot of a brown-white caravan and a horse on a meadow. The caravan is in the center, the horse in the back to its right, and there is a large tree on the right. The camera is slightly shaky, and there is a forested hill in the background.

Table 2: The rank of the search target using different models on VBS22 and VBS23 query sets. The post-fix "-t", "-r", "-u", and "-v3" denote the original query, the GPT-rephrased query, the user query, and the mean pooling of three visual queries, respectively. In the "-u" setting, a user is asked to find the target in the real KIS setting and the * indicates that the user found the search target.

	VBS22											
	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12
CLIP4Clipt-t	82	190	**1**	**1**	1966	12	4731	4	**2**	67747	261	**6**
CLIP4Clipt-t-r	2455	62	2	**1**	3876	**1**	4259	**1**	**2**	33742	582	150
CLIP4Clipt-t-u	61*	1134	2*	2*	**567**	45	156	51	10*	**12612**	25	71
CLIP4Clip-v3	**38**	**6**	**1**	**1**	2240	515	**25**	11108	12	13638	**8**	17
Enhanced ITV-t	**1**	15	627	**1**	62	11	**1**	**1**	1359	1553	5	**2**
Enhanced ITV-t-r	**1**	**9**	462	**1**	**33**	**1**	51	8	2233	**137**	173	40
Enhanced ITV-t-u	**1***	934	524*	**1***	502	97	3	24	315*	218	**2**	24
Enhanced ITV-v3	33	136	**1**	5	449	4	1329	764	**3**	1251	9	**2**
BLIP-t	2	**30**	**1**	**1**	26	229	**1**	3	13	955	24	3
BLIP-t-r	**1**	34	**1**	**1**	**2**	81	592	51	27	1067	26	27
BLIP-t-u	**1***	507	**1***	98*	35	1513	12	**2**	14*	**168**	**1**	**1**
BLIP-v3	**1**	130	**1**	8	23	**3**	**1**	67	**1**	2380	39	**1**
	VBS23											
	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12
CLIP4Clipt-t	**1**	8770	26	**186**	2059	1927	**145**	6756	**23**	167	**8**	280
CLIP4Clipt-t-r	**1**	2272	484	282	4605	29	450	2574	127	169	10	**1**
CLIP4Clipt-t-u	**1***	254	309	193	**601***	**19**	1268	**78**	62*	**49***	58	25
CLIP4Clip-v3	96	**49**	**8**	1135	3391	931	1381	7324	26	86	148	36
Enhanced ITV-t	**1**	**1**	170	394	225	88783	2554	558	**3**	66	**4**	**1**
Enhanced ITV-t-r	4	**1**	**23**	3420	4116	169	2608	**323**	**3**	15	41	2
Enhanced ITV-t-u	42*	**1**	312	8938	**44***	**21**	**1881**	**323**	8*	**4***	73	2
Enhanced ITV-v3	6	14	1061	**204**	1208	19911	2047	19856	742	89	1996	5
BLIP-t	**1**	**1**	30	**545**	237	323	912	386	35	24	176	35
BLIP-t-r	2	**1**	53	5189	31253	111	**245**	454	**25**	93	**46**	**11**
BLIP-t-u	**1***	13	87	1382	**152***	**16**	2869	**227**	53*	**18***	502	51
BLIP-v3	96	49	**8**	1135	3391	931	1381	7324	26	86	148	36

tasks on VBS 23 tasks. In addition, the rephrased query enhances the search target's rank in 30% to 50% of VBS queries. Notably, the performance of queries VBS23-KIS-t6 and VBS23-KIS-t12 consistently witnessed improvement across all baselines. The original captions of these two queries describe multiple scenes. For example, the query VBS23-KIS-t6 starts with "A sequence of three shots" and the three shots are almost independent. Although the query VBS23-KIS-t12 is not explicitly separated into multiple shots, the involved objects and events are scattered across diverse scenes. The query re-formulation of LLMs can help the model to concentrate on a specific scene. Oppositely, the performance on the query VBS23-KIS t4 and t5 consistently drop after revising the query. Compared to the original text, the rephrased t4 misses the key information "blue bow", "white/blue flower bouquet" and "red jacket", which are necessary to identify the target from the massive amount of wedding video shots. Although LLM can

Table 3: The mAP performance comparison of different models on TRECVid AVS query sets from tv19 to tv23. The post-fix "-t" indicated the performance using the original textual query and the post-fix "-vx" the visual query using the mean pooling of "x" generated image(s). The term "Fused" refers to the fusion of "-t" and "-v3" with the weights in the parentheses.

	tv19	tv20	tv21	tv22	tv23
CLIP4Clip-t	0.144	0.153	0.178	0.157	0.138
CLIP4Clip-v1	0.095	0.082	0.118	0.094	0.055
CLIP4Clip-v3	0.106	0.108	0.142	0.104	0.077
Fused (0.1, 0.9)	**0.156**	**0.167**	**0.191**	**0.170**	**0.145**
Enhanced ITV-t	0.163	**0.359**	0.355	**0.282**	0.292
Enhanced ITV-v1	0.189	0.199	0.218	0.162	0.191
Enhanced ITV-v3	0.219	0.220	0.240	0.181	0.212
Fused (0.1, 0.9)	**0.257**	0.347	**0.371**	0.243	**0.298**
BLIP2-t	0.199	0.222	**0.262**	0.164	0.204
BLIP2-v1	0.137	0.183	0.159	0.130	0.137
BLIP2-v3	0.179	0.211	0.189	0.159	0.190
Fused (0.1, 0.9)	**0.202**	**0.232**	0.227	**0.174**	**0.210**

concisely summarize the query, they cannot properly identify the key information for the video search. Nevertheless, human intervention can potentially mitigate these shortcomings by manually selecting one from multiple generated queries.

Table 3 compares the mAP performance on the TRECVid AVS tasks. Similarly, the model with post-fix "-t" indicates the original textual query, while "-vx" denotes mean pooling utilizing "x" image(s) produced by stable diffusion. The term "Fused" refers to an ensemble of "-t" and "-v3" with the specified weights in the parentheses. Generally, the mAP performance with textual queries is superior to the visual ones. In addition, with the increase in visual query numbers, the mAP performance gradually gets better. A possible reason is that the mean pooling of multiple images can provide a more general and stable representation. Notably, the fusion of textual and visual queries delivers the optimal mAP performance on most of the query sets. Among the baselines, the Enhanced-ITV consistently outperforms others across all AVS query sets.

4 Conclusion

We have presented the major improvements to our interactive system. Specifically, our system is equipped with GPT-4 and stable diffusion, allowing users to rephrase and imagine a text query, and then select the suitable LLM-refined queries for retrieval. We also improved the embedding techniques of the VIREO search engine using the enhanced ITV feature which demonstrates its effectiveness on the TRECVid AVS tasks.

5 Acknowledgments

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

References

1. Berns, F., Rossetto, L., Schoeffmann, K., Beecks, C., Awad, G.: V3c1 dataset: An evaluation of content characteristics. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. p. 334–338. ICMR '19 (2019)
2. Heller, S., Gsteiger, V., Bailer, W., Gurrin, C., ór Jónsson, B., Loko, J., Leibetseder, A., Mejzlík, F., Peka, L., Rossetto, L., Schall, K., Schoeffmann, K., Schuldt, H., Spiess, F., Tran, L.D., Vadicamo, L., Veselý, P., Vrochidis, S., Wu, J.: Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th video browser showdown. International Journal of Multimedia Information Retrieval **11**, 1 – 18 (2022)
3. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. ArXiv **abs/2301.12597** (2023)
4. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (2022)
5. Loko, J., Veselý, P., Mejzlík, F., Kovalcík, G., Soucek, T., Rossetto, L., Schoeffmann, K., Bailer, W., Gurrin, C., Sauter, L., Song, J., Vrochidis, S., Wu, J., ór Jónsson, B.: Is the reign of interactive search eternal? findings from the video browser showdown 2020. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**, 1 – 26 (2021)
6. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. Neurocomputing **508**, 293–304 (2021)
7. Nguyen, P.A., Ngo, C.W.: Interactive search vs. automatic search. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**, 1 – 24 (2021)
8. OpenAI: GPT-4 technical report. CoRR **abs/2303.08774** (2023)
9. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685 (jun 2022)
11. Rossetto, L., Schoeffmann, K., Bernstein, A.: Insights on the v3c2 dataset. arXiv preprint arXiv:2105.01475 (2021)
12. Schall, K., Hezel, N., Jung, K., Barthel, K.U.: Vibro: Video browsing with semantic and visual image embeddings. In: MultiMedia Modeling - 29th International Conference. vol. 13833, pp. 665–670. Springer (2023)
13. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. ArXiv **abs/2302.13971** (2023)
14. Wu, J., Ngo, C.W., Chan, W.K., Hou, Z.: (un)likelihood training for interpretable embedding. In: ACM Transactions on Information Systems (2023)