VideoEase at VBS2025: An Interactive Video Retrieval System

Quang-Linh Tran[1], Binh Nguyen[2], Gareth J. F. Jones[1], and Cathal Gurrin[1]

[1] ADAPT Centre, School of Computing, Dublin City University, Dublin 9, Ireland
linh.tran3@mail.dcu.ie, {gareth.jones,cathal.gurrin}@dcu.ie
[2] University of Science, Vietnam National University, Ho Chi Minh City, Vietnam
ngtbinh@hcmus.edu.vn

Abstract. We present the VideoEase interactive video retrieval system, which we used to participate in VBS2025. This is the first time that VideoEase has taken part in the VBS challenge. VideoEase is built on the Milvus vector database, which supports vector searches on massive datasets with high-dimensional vectors. The CLIP, BLIP2, and OpenCLIP models play a crucial role in encoding keyframe images and queries into embeddings. A new user interface with simple yet effective components is also introduced. We experiment with VBS24 queries to evaluate the performance of the VideoEase system. Our results show that answer rank is improved by merging outputs from several models with appropriate weights.

Keywords: Video Retrieval · Interactive Retrieval System · Multi-modal Embedding.

1 Introduction

With the increasing popularity of video and short-video platforms such as YouTube, TikTok, and Instagram, the number of videos on the internet is growing exponentially, with thousands to millions of new uploads daily. The quality of videos, in terms of resolution and frame rate, is also seeing significant improvements. The rapid growth in video archives poses a major challenge for the efficient organization, browsing, and retrieving of this content. Developers of state-of-the-art systems to address these challenges are invited to participate in the Video Browser Showdown (VBS) [10, 5]. Tis has provided an annual competition focused on video browsing and retrieval since 2012. VBS provides an evaluation platform for researchers worldwide to develop and test video retrieval solutions. VBS2025 is being organized in conjunction with the MMM conference in Nara, Japan. VBS2025 marks a significant upgrade to larger datasets from last year. Specifically, the third shard of the V3C (V3C3) dataset has been added, expanding the V3C dataset [2, 16] to 28,450 videos, totalling 3,800 hours of video content and an estimated 4.7 TB of data. In addition, the Marine Video Kit (MVK) dataset [20] has been expanded to MVK 2.0 with more videos. The LapGynLHE dataset remains unchanged from last year's competition, featuring 75 videos with a total of 104.75 hours of video content. With this increase in video content, VBS2025 is more challenging than ever, demanding significant optimization in organizing and indexing video content for efficient and accurate retrieval.

VBS2025 consists of two sessions and three task types. In the expert session, system owners perform the search tasks, while in the novice session, volunteers from the conference audience participate. The three tasks are Known-Item Search (KIS), Ad-hoc Video Search (AVS), and Question Answering (QA). The KIS task requires participants to locate the correct video segments as quickly as possible. Queries for the KIS task can be either video clips or textual descriptions. There is a penalty for incorrect submissions, and the first team to submit the correct answer receives the highest score. In the AVS task, participants are provided with a textual query and must locate as many relevant video segments as possible. Unlike the KIS task, the AVS task does not focus on a single correct segment but allows multiple relevant clips to match the query (e.g., clips showing a man using a mobile phone). The QA task, which was introduced in VBS2024, involves finding correct textual answers to specific questions. This task is particularly challenging as it requires not only retrieving the correct video segments but also extracting the correct answer from them.

Thirteen teams participated in the 2022 edition of the VBS competition. VISIONE 5.0 [1] won the competition with its diverse search functionalities, supporting searches using textual queries, images, objects, and colour drawings. The PraK Tool [11] excelled in the visual KIS task during the expert session with its advanced data service architecture. Vibro [17], which was also the winner of VBS2022 and VBS2023, was the best performer in the textual KIS task in the expert session, using EVACLIP-ViT-E for text-based searches and MixedSwim for image-based searches after extensive experimentation. DiveXplore [18] won the best QA task in the novice session in VBS2024. This system integrates OpenCLIP trained on the LAION-2B dataset for both text and visual similarity searches and features an optimized user interface for novice users. The Vitrivr [3] and Vitrivr-VR [19] systems share the same Vitrivr stack, including the Cottontail database, Cineast retrieval engine, and feature extraction components. However, Vitrivr employs a minimalistic user interface, while Vitrivr-VR utilizes a virtual reality user interface. With the rise of large language models (LLMs) in various applications, several teams have incorporated LLMs into their video retrieval systems. Ma et al. [12] utilized LLMs to expand and diversify the semantics of queries. Meanwhile, the TalkSee [4] system leverages LLMs to generate questions, update queries, and conduct re-ranking. The VERGE [14] system, which participated in VBS2024, offers various search functions and filters, including visual similarity, semantic similarity, free-text search, and filters by the number of people, colours, and activities. Khan et al. [8] implemented user relevance feedback and conversational search in their Exquisitor system. Waseda Meisei SoftBank [6] participated in VBS for the first time in 2024, building on an improved system from TRECVID 2023. The VideoCLIP 2 [13] and ViewInsight [21] systems were designed with a user-friendly interface for novice users, enhancing usability with a triple-image representation for each segment, allowing

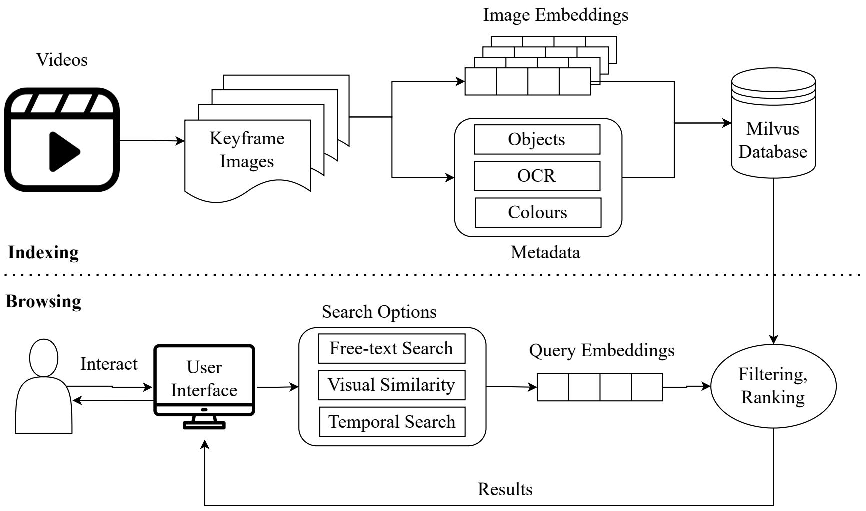


Fig. 1. The overview of VideoEase system.

users to easily grasp the meaning of each segment. From the previous systems in VBS2024, we observe the current approach is focusing on vision-language models for searching, and several search functions such as free-text search, image search, object and colour search, etc. We aim to leverage these approaches and enhance the user interface in our system.

This is the first time that VideoEase has taken part in the VBS competition, so we bring an entirely new system with several distinctions from previous participants. As a newcomer, we introduce a completely new video search engine and user interface. We use the Milvus[3] vector database to store metadata and vector embeddings of keyframe images in the dataset. Milvus is an open-source vector database designed specifically for similarity searches on massive datasets of high-dimensional vectors, making it highly suitable for dense vector retrieval in the VideoEase system. This database also provides high performance in search latency and accuracy compared to FAISS, which is used by most existing systems. We employ BLIP2 [9], CLIP [15], and OpenCLIP [7] for extracting image/text embeddings used in free-text and visual similarity searches. Finally, we have designed a new, easy-to-use interface tailored for novice users.

2 VideoEase System Overview

In this section, we provide an overview of the VideoEase system, along with a detailed explanation of its search functions. Figure 1 illustrates the system's archi-

[3] https://milvus.io/

tecture. The system operates in two stages: indexing (offline) and browsing (online). In the indexing stage, the videos are first split into keyframe images. These images are also provided by the organizers, including 4,143,683 images from the V3C dataset and 138,662 images from the LapGynLHE dataset. The keyframe images are processed by three pre-trained models: CLIP-ViT-L/14@336px [15], BLIP2-pretrain_vitL [9], and OpenCLIP-ViT-L-14 [7] (trained on the LAION2B dataset). The models generate embeddings from the images, which, along with metadata such as start and end times, objects, OCR (optical character recognition), and colours, are indexed in a Milvus collection.

In the browsing stage, users interact with the system's interface to search and browse keyframe images and videos. There are three search options available: free-text search, visual similarity search, and temporal search. Details about these options are provided in the following sections. User queries are transformed into embeddings using the same three pre-trained models. Milvus supports hybrid searches with multiple vectors and filtering in a single process, returning a list of keyframe images and associated videos. The results are displayed to users for further browsing and refinement of searches.

2.1 Free-text search with filters

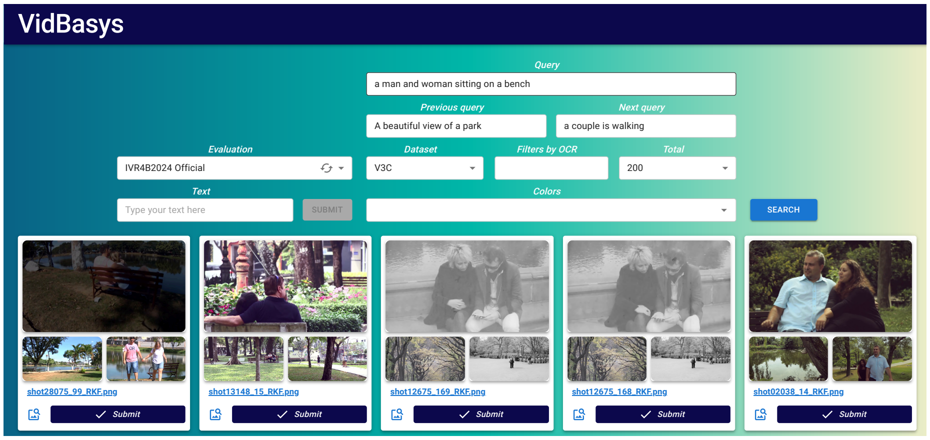
In the free-text search function, users enter a query into the search box. The query is a short text describing the content of a video clip. They can also apply filters, such as objects, colour, and free-text OCR from a drop-down list of options. These filters are used in the vector search process in Milvus as metadata filters. The query is then embedded into three vectors using the pre-trained models. A hybrid search request is sent to Milvus, which searches across all three embedding vectors with the applied filters. A reranking module combines the results from the three searches, producing a final ranked list. This module uses a weighted system, where each vector from the models is assigned a specific weight. The final results are displayed to users through the interface.

2.2 Visual similarity search

In the visual similarity search function, users can upload an image to search for visually similar images. Additionally, when results from a free-text search are displayed, users can select any of the images to initiate a visual similarity search. The selected images are embedded into vectors using the same three pre-trained models, and a hybrid search is performed, similar to the free-text search process. This search type is particularly useful for the AVS task, where users aim to find as many correct segments as possible for a given query. Users can locate a relevant segment and use it as input to search for additional related segments.

2.3 Temporal search

If users need to search for two scenes within the search, they can utilize the temporal search function. For example, a temporal query might be: "A girl and



Query: a man and woman sitting on a bench

Previous query: A beautiful view of a park

Next query: a couple is walking

Evaluation: IVR4B2024 Official

Dataset: V3C

Filters by OCR:

Total: 200

Text: Type your text here

Colors:

shot28075_99_RKF.png

shot13148_15_RKF.png

shot12675_169_RKF.png

shot12675_168_RKF.png

shot02038_14_RKF.png

Fig. 2. The user interface of the VideoEase system.

a man run up a small hill" for the main scene, followed by "There is a flagpole with a Canadian flag on top" for the next scene. The system first searches for the main scene, and then uses the video ID and start time of that scene to filter the search for the next scene. The results from both scenes are then combined using a weighted method. The final results are displayed with two images side-by-side.

2.4 User Interface

Our goal is to design a user-friendly interface that minimizes the number of required search actions. An example of the user interface is shown in Figure 2. The interface consists of two interactive parts. At the top is the search bar, where users can input queries. There are two fields for free-text queries: one for a single search and another for temporal searches (to be shown in an updated figure). Alongside the query fields are filters, allowing users to select or input search filters. The "Search" button initiates the search once all queries and filters are set.

The lower part of the interface displays the search results in a grid format, with keyframe images shown. Each image includes its corresponding video ID displayed beneath it. Users can click on an image to watch the video starting from that keyframe. Additionally, they can search for visually similar images by clicking the Visual Similarity icon. Once users are ready to submit, they can send the keyframe ID to the evaluation platform by clicking the "Submit" button.

3 Experimental Verification

To evaluate the accuracy of the search in this system, we conduct experiments analyzing the rank of correct answers in the returned result list. This is an automatic evaluation, with no user interactions involved. We use a set of queries

Table 1. Reranking experiment results. -1 indicates the system cannot find the answer in the top 10000. Bold text shows the best rank per query.

Reranking Weight (CLIP-BLIP2-OpenCLIP)	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12
0.3-0.3-0.4	4	1270	**1**	1795	**34**	**1**	4	1270	**1**	1795	**34**	**1**
1-0-0	**3**	2934	**1**	-1	185	4	**3**	2934	**1**	-1	185	4
0-1-0	67	1290	81	9563	843	6	67	1290	81	9563	843	6
0-0-1	**3**	**869**	**1**	**81**	167	**1**	**3**	**869**	**1**	**81**	167	**1**
RRFRanker	5	2323	**1**	275	164	**1**	5	2323	**1**	275	164	**1**

from the VBS2024 KIS task for this experiment, testing various weights assigned to each pre-trained model in the reranking process. Each weight reflects the influence of the cosine similarity score from a particular model on the final ranking list. Additionally, we apply the Reciprocal Rank Fusion (RRF) method for further comparison. The results are presented in Table 1 below.

The results from Table 1 indicate that incorporating all three models improves the ranking of answers for some queries. Overall, the system can find the correct answers within the top 50 results for 8 out of 12 queries, with 4 queries having the correct answer ranked first. OpenCLIP performs the best, identifying the most correct answers with the highest ranks, followed by CLIP and BLIP2. While the RRF method shows competitive performance on some queries, it still falls short of the best results.

4 Conclusions

In this paper, we introduce the VideoEase video retrieval system, which is participating in VBS2025 for the first time. The system supports several search functions and features an intuitive user interface designed for novice users. Our experiments demonstrate solid performance, with correct answers for 8 out of 12 queries appearing in the top 50 results.

Acknowledgement

This research was conducted with the financial support of Science Foundation Ireland at ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology at Dublin City University [13/RC/21-06_P2]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Amato, G., Bolettieri, P., Carrara, F., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: Visione 5.0: Enhanced user interface and ai models for vbs2024. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 332–339. Springer-Verlag, Berlin, Heidelberg (2024)
2. 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, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3323873.3325051, https://doi.org/10.1145/3323873.3325051
3. Gasser, R., Arnold, R., Faber, F., Schuldt, H., Waltenspül, R., Rossetto, L.: A new retrieval engine for vitrivr. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 324–331. Springer-Verlag, Berlin, Heidelberg (2024)
4. Gu, G., Wu, Z., He, J., Song, L., Wang, Z., Liang, C.: Talksee: Interactive video retrieval engine using large language model. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 387–393. Springer-Verlag, Berlin, Heidelberg (2024)
5. Heller, S., Gsteiger, V., Bailer, W., Gurrin, C., Jónsson, B.Þ., Lokoč, J., Leibetseder, A., Mejzlík, F., Peška, L., Rossetto, L., et al.: 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), 1–18 (2022)
6. Hori, T., Ueki, K., Suzuki, Y., Takushima, H., Tanoue, H., Sato, H., Takada, T., Kumar, A.M.: Waseda_meisei_softbank at video browser showdown 2024. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 311–316. Springer-Verlag, Berlin, Heidelberg (2024)
7. Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Open clip (7 2021)
8. Khan, O.S., Zhu, H., Sharma, U., Kanoulas, E., Rudinac, S., Jónsson, B.T.: Exquisitor at the video browser showdown 2024: Relevance feedback meets conversational search. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 347–355. Springer-Verlag, Berlin, Heidelberg (2024)
9. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023), https://arxiv.org/abs/2301.12597
10. Lokoč, J., Andreadis, S., Bailer, W., Duane, A., Gurrin, C., Ma, Z., Messina, N., Nguyen, T.N., Peška, L., Rossetto, L., Sauter, L., Schall, K., Schoeffmann, K., Khan, O.S., Spiess, F., Vadicamo, L., Vrochidis, S.: Interactive video retrieval in the age of effective joint embedding deep models: lessons from the 11th vbs. Multimedia Syst. **29**(6), 3481–3504 (aug 2023). https://doi.org/10.1007/s00530-023-01143-5, https://doi.org/10.1007/s00530-023-01143-5

11. Lokoč, J., Vopálková, Z., Stroh, M., Buchmueller, R., Schlegel, U.: Prak tool: An interactive search tool based on video data services. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 340–346. Springer-Verlag, Berlin, Heidelberg (2024)
12. Ma, Z., Wu, J., Ngo, C.W.: Leveraging llms and generative models for interactive known-item video search. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 380–386. Springer-Verlag, Berlin, Heidelberg (2024)
13. Nguyen, T.N., Quang, L.M., Healy, G., Nguyen, B.T., Gurrin, C.: Videoclip 2.0: An interactive clip-based video retrieval system for novice users at vbs2024. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 394–399. Springer-Verlag, Berlin, Heidelberg (2024)
14. Pantelidis, N., Pegia, M., Galanopoulos, D., Apostolidis, K., Stavrothanasopoulos, K., Moumtzidou, A., Gkountakos, K., Gialampoukidis, I., Vrochidis, S., Mezaris, V., Kompatsiaris, I., Jónsson, B.T.: Verge in vbs 2024. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 356–363. Springer-Verlag, Berlin, Heidelberg (2024)
15. 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 (2021), https://arxiv.org/abs/2103.00020
16. Rossetto, L., Schoeffmann, K., Bernstein, A.: Insights on the v3c2 dataset (2021), https://arxiv.org/abs/2105.01475
17. Schall, K., Hezel, N., Barthel, K.U., Jung, K.: Optimizing the interactive video retrieval tool vibro for the video browser showdown 2024. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 364–371. Springer-Verlag, Berlin, Heidelberg (2024)
18. Schoeffmann, K., Nasirihaghighi, S.: Divexplore at the video browser showdown 2024. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 372–379. Springer-Verlag, Berlin, Heidelberg (2024)
19. Spiess, F., Rossetto, L., Schuldt, H.: Exploring multimedia vector spaces with vitrivr-vr. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 317–323. Springer-Verlag, Berlin, Heidelberg (2024)
20. Truong, Q.T., Vu, T.A., Ha, T.S., Jakub, L., Tim, Y.H.W., Joneja, A., Yeung, S.K.: Marine video kit: A new marine video dataset for content-based analysis and retrieval (2022), https://arxiv.org/abs/2209.11518
21. Vuong, G.H., Ho, V.S., Nguyen-Dang, T.T., Thai, X.D., Le, T.K., Pham, M.K., Ninh, V.T., Gurrin, C., Tran, M.T.: Viewsinsight: Enhancing video retrieval for vbs 2024 with a user-friendly interaction mechanism. In: MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV. p. 400–406. Springer-Verlag, Berlin, Heidelberg (2024)