Fusionista2.0: Efficiency Retrieval System for Large-Scale Datasets

Huy M. Le[1,2,3,5], Dat Tien Nguyen[1,2], Phuc Binh Nguyen[3,5], Gia Bao Le Tran[3,5], Phu Truong Thien[3,5], Cuong Dinh[3,5], Minh Nguyen[3,5], Nga Nguyen[3,5], Thuy T. N. Nguyen[3,5], Huy Gia Ngo[3,5], Tan Nhat Nguyen[3,5], Binh T. Nguyen[2,4,5], and Monojit Choudhury[1]

[1] Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE
[2] AISIA Research Lab, Ho Chi Minh City
[3] University of Information Technology, Ho Chi Minh City, Vietnam
[4] Ho Chi Minh University of Science, Vietnam
[5] Vietnam National University, Ho chi Minh City, Vietnam
HuyM.Le@mbzuai.ac.ae

Abstract. The Video Browser Showdown (VBS) challenges systems to deliver accurate results under strict time constraints. To meet this demand, we present `Fusionista2.0`, a streamlined video retrieval system optimized for speed and usability. All core modules were re-engineered for efficiency: preprocessing now relies on ffmpeg for fast keyframe extraction, Optical Character Recognition (OCR) is powered by Vintern-1B-v3.5 for robust multilingual text recognition, and Automatic Speech Recognition (ASR) employs faster-whisper for real-time transcription. For question answering, lightweight vision–language models provide quick responses without the heavy cost of large models. Beyond these technical upgrades, `Fusionista2.0` introduces a redesigned UI/UX with improved responsiveness, accessibility, and workflow efficiency, enabling even non-expert users to retrieve relevant content rapidly. Evaluations demonstrate that retrieval time was reduced by up to 75% while accuracy and user satisfaction both increased, confirming `Fusionista2.0` as a competitive and user-friendly system for large-scale video search.

Keywords: Multimedia Retrieval · Efficient System · Multimodal.

1 Introduction

As noted in prior works [12], effective information retrieval not only improves user experience but also supports decision-making processes across a wide range of domains. Consequently, research efforts must focus on developing retrieval methodologies with superior efficacy or refining and optimizing existing approaches to enhance their scalability within this domain and foster seamless interaction between the retrieval system and the user. Addressing these points, Video Browser Showdown [19], is an established, biennial competition that serves as a robust benchmark for assessing state-of-the-art interactive video retrieval systems. The

VBS operates within realistic constraints, encompassing a spectrum of complex ad-hoc search issues, specifically: (*1*) Known-Item Search (KIS), (*2*) Ad-hoc Video Search (AVS), and (*3*) Visual Question Answering (VQA).

For VBS 2026, the primary dataset is the V3C [18], comprising over 28,000 videos with thousands of hours of content. Additional domain-specific datasets, such as Marine Video [23] and LapGynLHE [14], are also included. These heterogeneous datasets pose unique challenges and provide a comprehensive benchmark for testing the robustness and adaptability of retrieval systems.

In this paper, we present `Fusionista2.0`, an enhanced information retrieval system that extends `Fusionista` [9] and `Fustar` [8] with multiple modalities, including text-to-image, image-to-image, object-level, and OCR/ASR-based retrieval. The main improvements are:

(*a*) A redesigned UI/UX that enhances accessibility for non-expert users and reduces the learning curve.
(*b*) Optimized data processing and search mechanisms for faster retrieval.
(*c*) A powerful re-ranking module to refine results and maximize accuracy.

Statistics evaluations with new users show that `Fusionista2.0` have significantly improve in both efficiency and effectiveness, advancing interactive video retrieval while lowering the entry barrier for broader real-world use.

2 Key functions upgrade

2.1 Overview

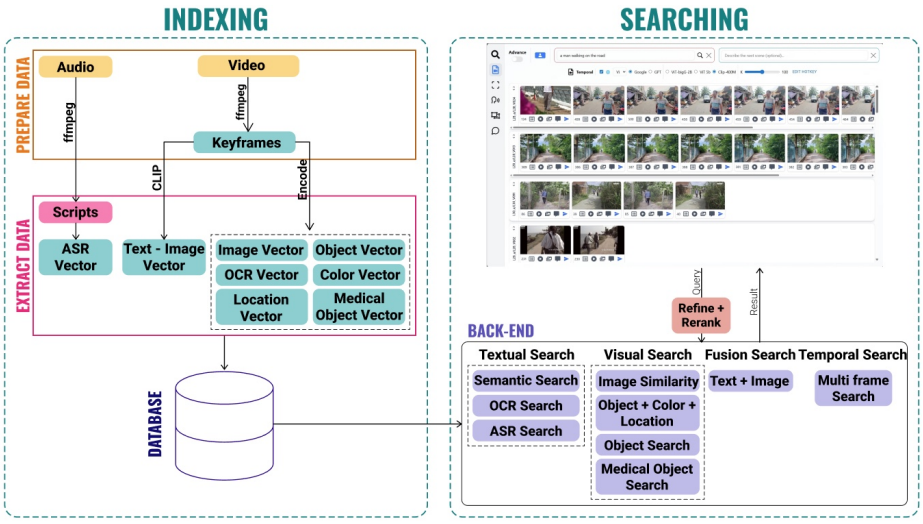

Fig. 1. Overview of out system.

Our system, `Fusionista2.0`, builds upon the design principles and techniques of its predecessors, `Fusionista` [9] and `Fustar` [8]. It provides a comprehensive set of retrieval functionalities tailored for the diverse query types in

VBS2025. While these functionalities proved effective in VBS2025, scaling the system to multi-terabyte datasets revealed several challenges such as inference time, UX flow for thousand images. In `Fusionista2.0`, we introduce six key improvements (Figure 1), including:

1. Optimized Data Preparation: The new all-in-one preprocessing pipeline is highly optimized for video decoding, frame extraction, and keyframe selection. This ensures faster, resource-efficient processing of large-scale datasets.
2. Enhanced Textual Search: Multiple embedding models are combined into an ensemble to improve semantic coverage and accuracy.
3. Efficient OCR/ASR: Large OCR and ASR models are replaced by lightweight alternatives with comparable accuracy, ensuring scalability.
4. Modern VLLM-based Question Answering: The QA module is upgraded with recent Vision–Language Large Models (VLLMs), optimized for both reasoning and inference speed, delivering faster and accurate answers.
5. Reranking with Interactive Confirmation: A novel reranking mechanism leverages clarification questions to refine the search results according to user intent, reducing the risk of missing subtle details.
6. UI/UX Enhanced: As user interaction is the most critical aspect in interactive retrieval, we redesigned the interface to minimize redundant browsing of repeated video content, aligning the layout with natural user viewing flow.

Through these improvements, `Fusionista2.0` achieves a balanced trade-off between retrieval effectiveness, efficiency, and user interactivity, addressing the challenges of large-scale multimedia search in VBS2026.

2.2 Data Preparation and Extraction

In our previous system, the preprocessing pipeline was designed to maximize the accuracy of keyframe selection by leveraging multiple state-of-the-art models. Our previous system employed a preprocessing pipeline that combined CLIP-B/32 embeddings [16], TransNetV2 [20] scene detection, and clustering to select representative keyframes [22], which were then used for multi-function search.

Although this method achieved high precision in selecting representative keyframes, it was computationally intensive, requiring significant GPU and memory resources, and thus proved inefficient for large-scale datasets like V3C.

To overcome scalability issues, the multi-model pipeline is replaced with an all-in-one *ffmpeg*-based keyframe extraction workflow. The system deterministically analyzes each video stream and isolates all intra-coded frames that serve as structural keyframes, exporting them as sequential images along with a precise timestamp–frame mapping for reproducibility. This streamlined approach greatly reduces computational and memory demands, enabling efficient large-scale preprocessing while maintaining sufficient quality for frame-extraction tasks, thus ensuring feasibility under the competition's constraints.

2.3 Textual Search

In our previous system, text-to-image retrieval relied on a single CLIP-based model, which often required a trade-off between efficiency and accuracy. In

`Fusionista2.0`, we enhance this module by employing two state-of-the-art CLIP variants [2]: *CLIP-Sig400M* and *CLIP-ViT-5B*. Both models are widely recognized for their balanced performance in terms of inference speed and retrieval accuracy, making them well-suited for large-scale interactive search scenarios.

$$s(q, v) = \alpha \cdot s_{\text{Sig400M}}(q, v) + (1 - \alpha) \cdot s_{\text{ViT-5B}}(q, v) \tag{1}$$

The value of α is fixed equal to 0.7, which is empirically chosen based on extensive internal experimentation and a small-scale user study (Table 1) conducted with 50 participants interacting with our system to find best result for a text-image retrieval question. This adaptive weighting allows `Fusionista2.0` to capture semantic nuances more robustly than using either model in isolation.

Table 1. Number of people can retrieve the result at top-1 with different α thresholds and model backbones on the person-matching task.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Number of People	34	35	31	40	39	38	43	29	38
CLIP Sig400M	38								
CLIP ViT-5B	33								

2.4 OCR and ASR

In `Fusionista2.0`, we replace the previous PaddleOCR pipeline with the *Vintern-1B-v3.5* model [3]. This model, fine-tuned from InternVL2.5-1B [1], leverages low-resource languages datasets such as Viet-ShareGPT-4o-Text-VQA [3] and over one million OCR samples, achieving strong performance in many languages. Vintern-1B-v3.5 not only delivers high accuracy in scene text detection and recognition but also benefits from enhanced reasoning and general knowledge, enabling it to infer characters that are blurred or occluded.

While Whisper [17] provides excellent accuracy, its large model size makes it impractical for large-scale processing. Moreover, the majority of audio tracks in the VBS datasets contain ambient sounds rather than speech, reducing the necessity for high-capacity ASR models. For these reasons, we replace Whisper with *faster_whisper* [21], a lightweight yet efficient variant optimized for speed. This choice significantly accelerates the transcription pipeline by four times [21], while still maintaining sufficient accuracy to extract meaningful speech content.

2.5 Question Answering

Table 2. Summary performance per model.

Model	Count Acc.	Img. Ans. Acc.	Video Anss Acc.	Avg. Time (s)
InternVL-1B-Seq [1]	0.72	0.68	0.67	2.52
InternVL-1B-ffn6 [1]	0.79	0.74	0.71	2.48
InternVL-1B-ffn6-Seq [1]	**0.84**	**0.79**	**0.75**	4.80
LLaVA-0.5B-ffn6 [10]	0.78	0.63	0.73	10.50
SmolVLM-0.5B-ffn6 [13]	0.83	0.52	0.65	**1.07**

While large vision-language models (VLMs, ≥7B) excel in video question answering, they are too slow and often less accurate than humans, making them impractical for time-critical retrieval tasks like VBS. Our approach prioritizes speed and reliability by targeting simple but frequent queries—such as object counting, attribute recognition, and text extraction—while offering partial support for human for complex reasoning tasks. This strategy ensures fast responsiveness while keeping human expertise in the loop where models still fall short.

To systematically choose the right lightweight models, we constructed a benchmark dataset of 200 video–question pairs from previous queries of VBS, categorized into three main groups: (1) Counting (*How many completed shoes in the image?*), (2) Image information extraction (*What is on the street?*), and (3) Video information extraction (*What color is the phone the woman is using?*). The dataset requires models to provide an image/video. We then tested several compact VLMs (≤1B parameters) on this dataset, with the aim of identifying the best compromise between accuracy and inference speed. The detailed results are summarized in Table 2.

Among all candidates, the *InternVL-1B-ffn6-Seq* [1] model demonstrated the most consistent performance across the three categories. It achieved an average inference time of less than 5 seconds per query over the entire dataset, while maintaining high accuracy in counting and frame-level information extraction tasks. This balance of speed and reliability makes it the most suitable choice for our system, ensuring accurate answers within the strict time limits of VBS.

2.6 Reranking

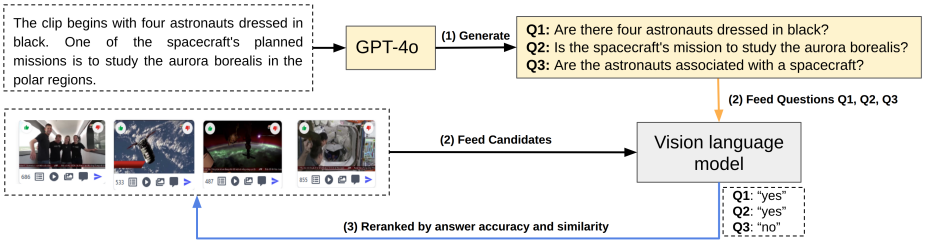

Fig. 2. Illustration of reranking process. Images from prior search is reranked based on number of yes answers from the Vision language model

Our reranking approach is inspired by the methodology outlined in [5, 7], which utilizes a post-processing strategy that can be effortlessly incorporated into existing image retrieval systems. However, we extend this concept by designing a fully integrated AI-based VQA pipeline. In this pipeline, we employ GPT-4o [15] to formulate three yes-no questions based on the initial user query. These questions are crafted to explore the details of objects mentioned in the query, such as "*Is there a dog in the scene?*" or "*Is the dog colored yellow?*" Subsequently, both the questions and the associated images are evaluated by a vision language model to generate accurate responses. To enhance flexibility, we test

multiple vision language models, including VideoLLaMA [27] and BLIP-2 [11], leveraging the VLLMs [6] library to enable API-driven interactions.

3 UI/UX Upgrade

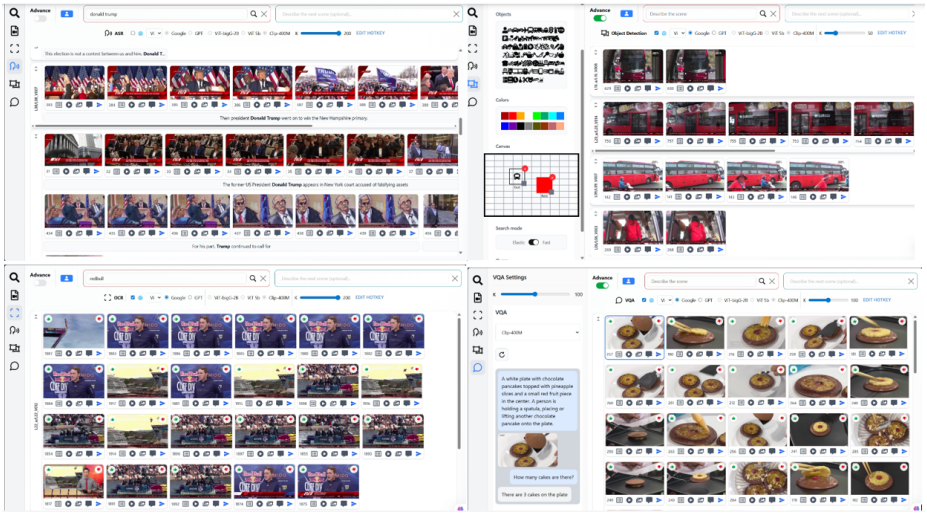

Fig. 3. Our system's UI/UX for main functions. On the top-left, top-right, bottom-left, bottom-right is ASR search, object search, OCR search, QA respectively.

Our main key focus is UI/UX optimization, as efficient interaction directly impacts retrieval speed and usability. We redesigned the interface for responsiveness, accessibility (*WCAG* [26] compliance with *shadcn/ui*) [24], better feedback via loading states and error handling. Migration from *Create React App (CRA)* [4] to *Vite* [25] enhance significantly system performance.

Workflows were streamlined with grouped search results, sidebar navigation with shortcuts, virtual scrolling, and multilingual query support. A conversational VQA interface and batch operations further boosted efficiency. These changes had a measurable impact on user interactions (showed in Figure 3. Overall, the optimizations greatly enhanced responsiveness, accuracy, and user satisfaction, ensuring `Fusionista2.0`'s competitiveness in time-critical video search.

4 Conclusion

`Fusionista2.0` demonstrates that optimizing speed across all modules and redesigning the UI/UX can significantly improve both efficiency and usability in large-scale video retrieval. These advances make the system highly competitive for VBS2026 while lowering the barrier for real-world adoption.

Acknowledgments. This research was supported by The Mohamed bin Zayed University of Artificial Intelligence's Scientific Research Fund and The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

1. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
2. Cui, Y., Zhao, L., Liang, F., Li, Y., Shao, J.: Democratizing contrastive language-image pre-training: A CLIP benchmark of data, model, and supervision. CoRR **abs/2203.05796** (2022). https://doi.org/10.48550/ARXIV.2203.05796
3. Doan, K.T., Huynh, B.G., Hoang, D.T., Pham, T.D., Pham, N.H., Nguyen, Q.T.M., Vo, B.Q., Hoang, S.N.: Vintern-1b: An efficient multimodal large language model for vietnamese (2024), https://arxiv.org/abs/2408.12480
4. Facebook: create-react-app: Set up a modern web app by running one command. https://github.com/facebook/create-react-app, accessed: 2025-09-17
5. Feng, C., Bai, Y., Luo, T., Li, Z., Khan, S.H., Zuo, W., Goh, R.S.M., Liu, Y.: VQA4CIR: boosting composed image retrieval with visual question answering. In: Walsh, T., Shah, J., Kolter, Z. (eds.) AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. pp. 2942–2950. AAAI Press (2025). https://doi.org/10.1609/AAAI.V39I3.32301
6. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (2023)
7. Le, H.M., Luong, V.T., Luong, N.H.: Data augmentation with large language models for vietnamese abstractive text summarization. In: International Conference on Multimedia Analysis and Pattern Recognition, MAPR 2023, Quy Nhon, Vietnam, October 5-6, 2023. pp. 1–6. IEEE (2023). https://doi.org/10.1109/MAPR59823.2023.10288906
8. Le, H.M., Tien, D.N., Duy, K.L., Quang, T.N.D., Khanh, T.N., Nguyen, T., Nguyen, B.T.: Fustar: Divide and Conquer Query in Video Retrieval System, Information and Communication Technology, vol. 2353. Springer Nature Singapore :, Singapore :, 1st ed. 2025. edn. (2025)
9. Le, H.M., Tien, D.N., Duy, K.L., Quang, T.N.D., Toan, N.K., Nguyen, T., Nguyen, B.T.: Fusionista: Fusion of 3-d information of video in retrieval system. In: MultiMedia Modeling: 31st International Conference on Multimedia Modeling, MMM 2025, Nara, Japan, January 8–10, 2025, Proceedings, Part V. p. 278–285. Springer-Verlag, Berlin, Heidelberg (2025). https://doi.org/10.1007/978-981-96-2074-6_33
10. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. Trans. Mach. Learn. Res. **2025** (2025), https://openreview.net/forum?id=zKv8qULV6n
11. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
12. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
13. Marafioti, A., Zohar, O., Farré, M., Noyan, M., Bakouch, E., Cuenca, P., Zakka, C., Allal, L.B., Lozhkov, A., Tazi, N., Srivastav, V., Lochner, J., Larcher, H., Morlon,

M., Tunstall, L., von Werra, L., Wolf, T.: Smolvlm: Redefining small and efficient multimodal models. arXiv preprint arXiv:2504.05299 (2025)
14. Nasirihaghighi, S., Ghamsarian, N., Peschek, L., Munari, M., Husslein, H., Sznitman, R., Schoeffmann, K.: Gynsurg: A comprehensive gynecology laparoscopic surgery dataset. CoRR **abs/2506.11356** (2025). https://doi.org/10.48550/ARXIV.2506.11356
15. OpenAI: Gpt-4o system card (2024), https://arxiv.org/abs/2410.21276
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of ICML (2021)
17. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. Proceedings of Machine Learning Research, vol. 202, pp. 28492–28518. PMLR (2023), https://proceedings.mlr.press/v202/radford23a.html
18. Rossetto, L., Lokoc, J., Bailer, W., Schoeffmann, K., Awad, G.: V3c–a research video collection. In: Proceedings of the 9th ACM Multimedia Systems Conference. pp. 461–466 (2019)
19. Schoeffmann, K., Boeszoermenyi, L., Beesley, P.: The video browser showdown: a live evaluation of interactive video retrieval systems. International Journal of Multimedia Information Retrieval **1**(3), 207–227 (2012)
20. Souček, T., Lokoč, J.: Transnet v2: An effective deep network architecture for fast shot transition detection. In: Proceedings of the IEEE International Conference on Image Processing (ICIP). pp. 151–155 (2020)
21. SYSTRAN: faster-whisper: Transcription with ctranslate2. https://github.com/SYSTRAN/faster-whisper (2024), gitHub repository, MIT License
22. Tan, K., Zhou, Y., Xia, Q., Liu, R., Chen, Y.: Large model based sequential keyframe extraction for video summarization. In: Proceedings of the International Conference on Computing, Machine Learning and Data Science, CMLDS 2024, Singapore, April 12-14, 2024. pp. 52:1–52:5. ACM (2024). https://doi.org/10.1145/3661725.3661781
23. Truong, Q., Vu, T., Ha, T., Lokoc, J., Wong, Y.H.T., Joneja, A., Yeung, S.: Marine video kit: A new marine video dataset for content-based analysis and retrieval. In: MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023, Proceedings, Part I. Lecture Notes in Computer Science, vol. 13833, pp. 539–550. Springer (2023). https://doi.org/10.1007/978-3-031-27077-2_42
24. shadcn ui: shadcn-ui/ui: A set of beautifully-designed, accessible components and a code distribution platform. https://github.com/shadcn-ui/ui, accessed: 2025-09-17
25. Vite contributors: Guide — vite. https://vite.dev/guide/, accessed: 2025-09-17
26. World Wide Web Consortium (W3C): Web content accessibility guidelines (wcag) 2.1. https://www.w3.org/TR/WCAG21/ (2018), accessed: 2025-09-17
27. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023. pp. 543–553. Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.EMNLP-DEMO.49