DataCube: A Video Retrieval Platform via Natural Language Semantic Profiling

Yiming Ju , **Hanyu Zhao** , **Quanyue Ma** , **Donglin Hao** , **Chengwei Wu** , **Ming Li** , **Songjing Wang** and **Tengfei Pan**

Beijing Academy of Artificial Intelligence

{ymju, hyzhao, tfpan}@baai.ac.cn

Abstract

Large-scale video repositories are increasingly available for modern video understanding and generation tasks. However, transforming raw videos into high-quality, task-specific datasets remains costly and inefficient. We present DataCube, an intelligent platform for automatic video processing, multi-dimensional profiling, and query-driven retrieval. DataCube constructs structured semantic representations of video clips and supports hybrid retrieval with neural re-ranking and deep semantic matching. Through an interactive web interface, users can efficiently construct customized video subsets from massive repositories for training, analysis, and evaluation, and build searchable systems over their own private video collections. The system is publicly accessible at *https://datacube.baai.ac.cn/*. **Demo Video:** *https://youtu.be/L7bKfPBm2tU*

1 Introduction

The rapid development of video generation and understanding models has raised higher demands on data quality, composition, and customization [Zhou *et al.*, 2024]. Although large-scale open-domain video repositories and open-source datasets at the petabyte scale are increasingly available, such as HowTo100M [Miech *et al.*, 2019], Panda-70M [Chen *et al.*, 2024b], Koala-36M [Wang *et al.*, 2025], and several recent large-scale video datasets [Nan *et al.*, 2024; Ju *et al.*, 2025a; Wang *et al.*, 2023; Chen *et al.*, 2024a; Ju *et al.*, 2025b]. Most of them cannot be directly utilized due to the difficulty of extracting task-specific content and controlling dataset distributions. As a result, constructing customized video datasets remains costly and time-consuming.

Previous works have explored CLIP-based embedding methods for indexing large-scale video repositories and enabling similarity-based retrieval [Luo *et al.*, 2022; Portillo-Quintero *et al.*, 2021; Nguyen *et al.*, 2024; Zhao *et al.*, 2022]. While effective and efficient in practice, these approaches typically rely on a single global embedding to capture overall visual–textual similarity. Such representations offer limited flexibility for modeling fine-grained semantics and make it difficult to extend the system once the indexing pipeline is established.

Recent advances in multimodal understanding models have made it possible to transform video content into structured natural-language representations [Yin *et al.*, 2024]. Based on such representations, videos can be described as collections of interpretable semantic attributes, enabling semantic retrieval and analysis. Compared with CLIP-based similarity retrieval methods, this paradigm is more scalable, as semantic annotations can be flexibly extended to multiple attributes, and more interpretable, facilitating error diagnosis and semantic tracing. By leveraging pre-computed semantic representations over massive video repositories, such semantic-aware retrieval systems enable large-scale video understanding and personalized data access. Users can efficiently construct customized datasets without repeatedly reprocessing the entire corpus, thereby avoiding redundant computation and unnecessary resource consumption.

In this work, we present DataCube, an end-to-end intelligent platform for video processing, profiling, and semantic retrieval. DataCube integrates automatic preprocessing, multi-dimensional semantic modeling, and hybrid retrieval mechanisms to enable efficient construction of customized video datasets from large-scale repositories. By transforming raw videos into structured natural-language semantic representations, DataCube supports flexible and personalized video retrieval based on content, style, and other semantic constraints. A coarse-to-fine retrieval framework is adopted to balance efficiency and accuracy, combining embedding-based filtering with deep semantic matching for complex queries.

Built upon large-scale open datasets and continuously expanded through user-contributed repositories, DataCube maintains a searchable index of hundreds of millions of video clips. The platform is publicly accessible through a web interface at https://datacube.baai.ac.cn/, supports both Chinese and English queries, and has attracted over one thousand registered users. In addition, it has been adopted in collaborative research projects with multiple universities.

2 System Overview

DataCube is an end-to-end platform for automatic video processing, profiling, and semantic retrieval, which is built upon large-scale open datasets, including Runway and

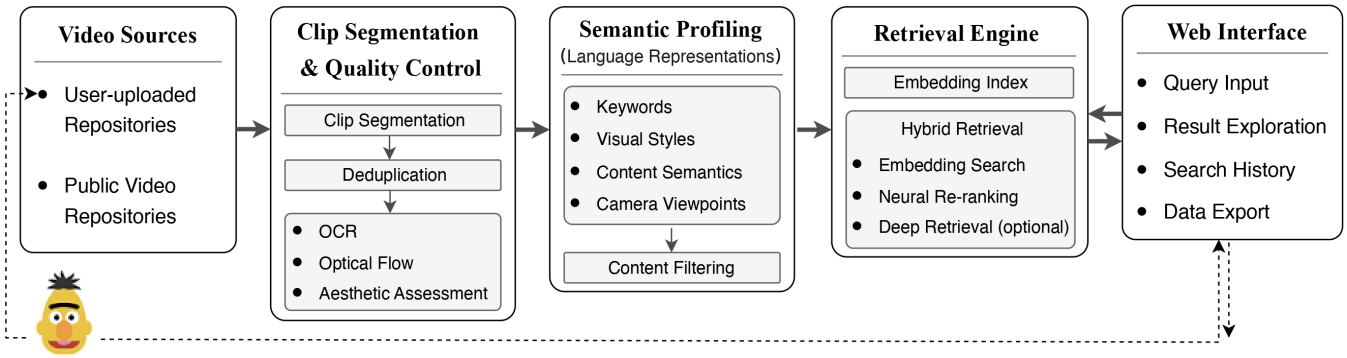

Figure 1: System Overview.

Koala [Wang *et al.*, 2025] collections, and supports data ingestion from user-uploaded private repositories. As illustrated in Figure 1, the system transforms large-scale raw video repositories into profile-enriched and searchable databases through a unified pipeline, enabling customized retrieval and dataset construction.

As shown in Figure 1, video sources are first automatically segmented into clips and processed by a multi-stage quality control pipeline, including deduplication, optical flow analysis, OCR detection, and aesthetic assessment. Next, DataCube constructs multi-dimensional semantic profiles in natural language using large vision-language models (VLMs). These profiles capture visual styles, content semantics, camera viewpoints, and associated keywords, and are further used to filter potentially non-compliant videos. The remaining profiles are then encoded into dense embeddings for efficient indexing.

On the retrieval side, user queries are processed by a hybrid retrieval engine. Queries are first enriched and analyzed, and then mapped to different profiling dimensions. The mapped representations are used for embedding-based retrieval and neural re-ranking. A deep retrieval mode is further provided for complex queries, enabling direct video-query comparison for high-precision filtering. Finally, DataCube exposes all functionalities through an interactive web interface, allowing users to explore retrieval results, browse historical search records, and export customized video subsets.

3 Key Components and Implementation

This section outlines the key implementation of DataCube.

Data Storage and Preprocessing. All videos and generated clips are stored in KS3 object storage. Raw videos are segmented into semantic clips using content-aware scene detection with PySceneDetect[1] (threshold = 26). Clips shorter than five seconds are discarded, while raw videos are archived in cold storage. During preprocessing, DataCube applies URL-based and frame-hash-based deduplication. Text coverage is estimated using PaddleOCR[2], motion strength is measured by computing optical flow every 0.5 seconds with RAFT [Teed and Deng, 2020], and clip-level aesthetic quality is estimated using NIQE [Mittal *et al.*, 2013] and MUSIQ [Ke *et al.*, 2021]. Based on these signals, nearly static clips (motion score < 10) are filtered out, while other attributes are retained as user-configurable filtering options.

Semantic Profiling and Indexing. Qualified clips are analyzed using Qwen2.5-VL-7B [Wang *et al.*, 2024] to construct natural-language semantic profiles, covering keywords, visual styles, content semantics, and camera viewpoints. Each clip is represented by a concatenated image composed of uniformly sampled frames. The resulting profiles are encoded into dense embeddings using BGE[3] and indexed with Milvus. CPU and GPU workloads are dynamically scheduled using Ray, enabling scalable indexing over hundreds of millions of clips.

Hybrid and Deep Retrieval. User queries are first enriched using a GPT-based interface and encoded with BGE embeddings to retrieve candidate subsets from Milvus (default size: 10,000). The retrieved candidates are then reranked using Qwen3-Reranker-0.6B [Zhang *et al.*, 2025]. For high-precision requirements, Qwen2.5-VL-72B [Wang *et al.*, 2024] performs direct video-query semantic matching for deep retrieval. The deep retrieval module is deployed with vLLM [Kwon, 2025] on A100 GPUs and supports long-running background inference tasks.

4 Demonstration Scenario

DataCube exposes all functionalities through an interactive web interface, allowing users to efficiently explore retrieval results, browse historical search records, and export customized video subsets. Figure 2 illustrates the overall demonstration workflow.

Interactive Query and Retrieval. Users can **select a data source** for retrieval, either from the public platform repository or from private repositories[4]. When creating a repository, users may choose whether to **share it with other users**

[1]https://github.com/Breakthrough/PySceneDetect
[2]https://github.com/PaddlePaddle/PaddleOCR
[3]https://huggingface.co/BAAI/bge-large-en-v1.5
[4]Users can upload raw videos to DataCube, which automatically processes the data and constructs an indexed repository. Once processing is completed, the repository becomes selectable for retrieval. DataCube supports retrieval from multiple data sources, which can be freely combined within a single query.

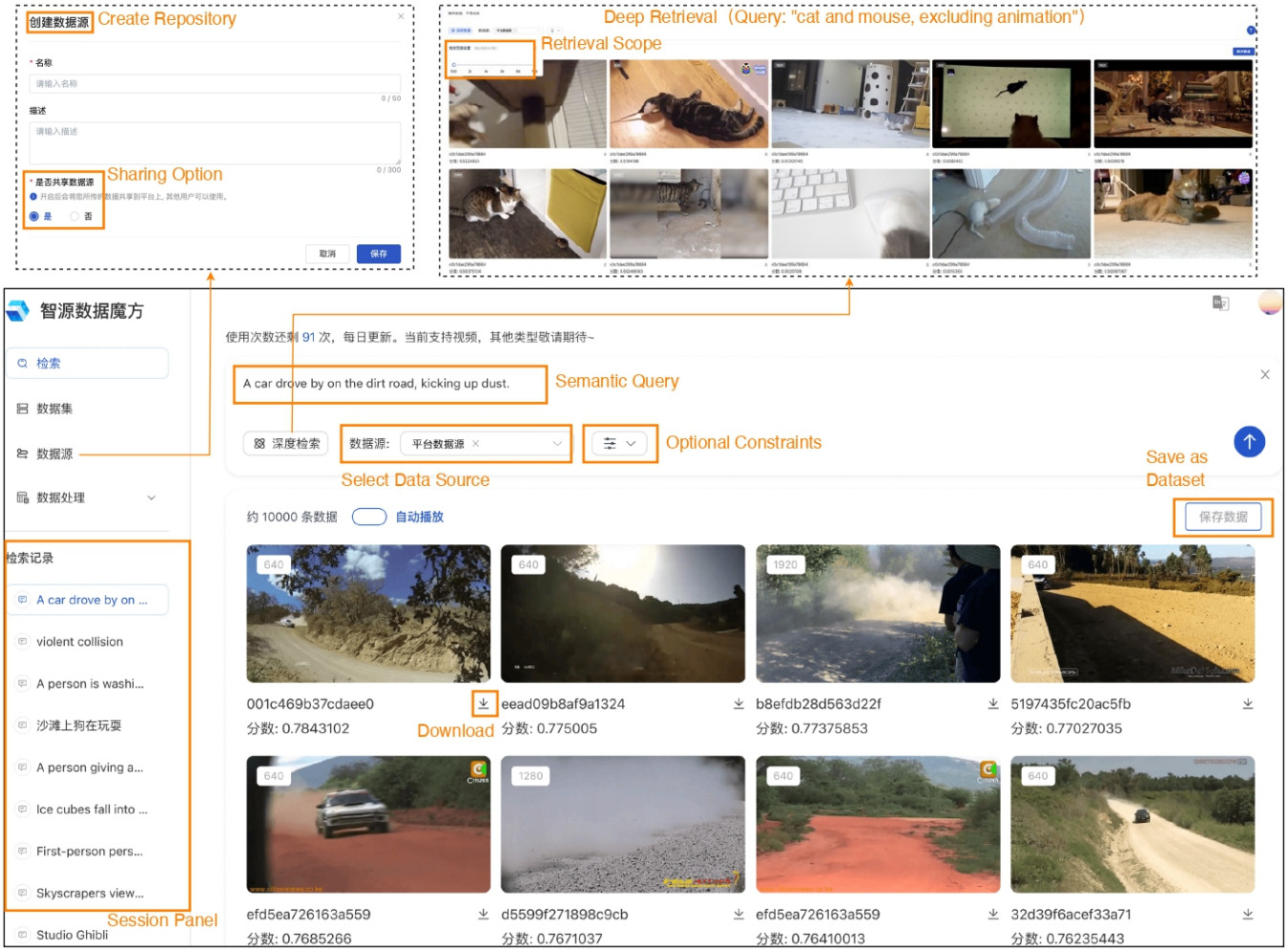

Figure 2: Demonstration Scenario of DataCube.

or keep it private. During retrieval, users specify **semantic queries** in natural language and configure **optional constraints** through a control panel, such as target resolution and clip duration ranges. After submission, DataCube executes the hybrid retrieval pipeline and presents ranked results in a grid layout, where each video is displayed together with its relevance score. Retrieved results are simultaneously recorded in a **session panel** on the left sidebar, enabling users to track multiple retrieval tasks.

Result Management and Dataset Export. Individual videos can be **previewed and downloaded** directly through the web interface. Users may further select a subset of the top-ranked retrieved videos (up to 10,000) and **save them as a dataset**, upon which the system automatically packages the selected samples and generates a downloadable link for offline usage.

Deep Retrieval for Complex Queries. Users can enable the **deep retrieval** mode. In this mode, DataCube launches background matching tasks and records their execution status in the sidebar. Due to the increased computational cost, this mode processes up to 500 candidate videos by default and typically completes within 3–5 minutes. Users may optionally expand the **retrieval scope** to 10,000 candidates, with processing time increasing approximately linearly. Compared with standard retrieval, deep retrieval provides more accurate semantic matching and supports complex constraints. For example, Figure 2 presents retrieval results for a query that includes an explicit exclusion requirement. Once the task is completed, refined results can be accessed from the sidebar and downloaded as a filtered dataset.

5 Conclusion

We introduce DataCube, an end-to-end platform for large-scale video processing and semantic retrieval. By combining structured semantic profiling with hybrid and deep retrieval mechanisms, DataCube enables efficient construction of customized video datasets from massive repositories through an interactive web interface. The system demonstrates the practicality of semantic-aware video retrieval by promoting data sharing and semantic profile reuse, thereby effectively reducing data preparation costs and supporting video-centric research and applications.

Contribution Statement

Yiming Ju and Hanyu Zhao contributed equally to this work. Tengfei Pan is the corresponding author.

References

[Chen *et al.*, 2024a] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024.

[Chen *et al.*, 2024b] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024.

[Ju *et al.*, 2025a] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2025.

[Ju *et al.*, 2025b] Yiming Ju, Jijin Hu, Zhengxiong Luo, Haoge Deng, hanyu Zhao, Li Du, Chengwei Wu, Donglin Hao, Xinlong Wang, and Tengfei Pan. Ci-vid: A coherent interleaved text-video dataset, 2025.

[Ke *et al.*, 2021] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5148–5157, 2021.

[Kwon, 2025] Woosuk Kwon. *vLLM: An Efficient Inference Engine for Large Language Models*. PhD thesis, UC Berkeley, 2025.

[Luo *et al.*, 2022] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

[Miech *et al.*, 2019] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[Mittal *et al.*, 2013] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[Nan *et al.*, 2024] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *ArXiv preprint*, abs/2407.02371, 2024.

[Nguyen *et al.*, 2024] Thao-Nhu Nguyen, Le Minh Quang, Graham Healy, Binh T Nguyen, and Cathal Gurrin. Videoclip 2.0: an interactive clip-based video retrieval system for novice users at vbs2024. In *International Conference on Multimedia Modeling*, pages 394–399. Springer, 2024.

[Portillo-Quintero *et al.*, 2021] Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12. Springer, 2021.

[Teed and Deng, 2020] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[Wang *et al.*, 2023] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *ArXiv preprint*, abs/2307.06942, 2023.

[Wang *et al.*, 2024] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Wang *et al.*, 2025] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437, 2025.

[Yin *et al.*, 2024] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

[Zhang *et al.*, 2025] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[Zhao *et al.*, 2022] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 970–981, 2022.

[Zhou *et al.*, 2024] Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. A survey on generative ai and llm for video generation, understanding, and streaming. *arXiv preprint arXiv:2404.16038*, 2024.