Article

Not peer-reviewed version

A Cross-modal Retrieval Method for Image, Audio and Video Based on OpenKylin

Jin Zhang [†] , Xin Xie [†] , Xiao dong Liu [*] , Jie Yu , Long Peng [†] , Wen zhu Wang [†] , Peng fei Zhang [†] , Yu Lan [†] , Chao Zhang [†]

Posted Date: 3 November 2023

doi: 10.20944/preprints202311.0185.v1

Keywords: cross-modal retrieval; open Kylin; domestic operating system

Article

A Cross-modal Retrieval Method for Image, Audio and Video Based on OpenKylin

Jin Zhang1,‡ , Xin Xie1,‡ , Xiaodong Liu2, *,Jie Yu2,‡, Long Peng [2,‡], Wenzhu Wang [3,‡], Pengfei Zhang [4,‡], Yue Lan [5,‡] and Chao Zhang [5,‡]

1 College of Computer and Communication Engineering Changsha University of Science and Technology, Changsha 41005,China; mail_zhangjin@163.com
2 School of Computer Science National University of Defense Technology, Changsha 41003, China
3 Haihe Laboratory of Information Technology Application Innovation, Tianjin 300462, China
4 Kylin Software, Tianjin 300462, China
5 Kylin Software, Changsha 410153, China
* Correspondence: maxxie622@163.com;
‡ These authors contributed equally to this work.

Abstract: The need for cross-modal retrieval (CMR) among users is growing because of the swift advancements in computer vision and natural language processing. Promoting home operating systems is greatly aided by the openKylin home operating system. It is focused on enhancing the functionality of the operating system with safe and dependable operating technologies. Practical implications abound for integrating computer vision research into home operating systems. A multimodal retrieval approach based on OpenKylin is presented to improve the usability of text search graphs, audio, and video. The overview of pertinent domestic and international research is the main topic of this article. This technological advancement. It then offers details on the design, implementation procedure, and final implementation outcomes for demonstration. The works of natural operations show that this strategy has excellent accuracy and performance.

Keywords: cross-modal retrieval; open Kylin; domestic operating system

1. Introduction

Based on the rapid development of the Internet and the rapid popularization of social media, a vast amount of data and information in different modalities have been generated, and people can conveniently access the vast amount of data from the new media technology. Each form of data can be viewed as a modality, such as text, image, video, etc. Cross-modal retrieval[1–4]explores the relationship between different modal samples,i.e., using one modal data as a query condition to retrieve the related data of another modality.

Natural language processing and computer vision are currently two of the hottest research directions in artificial intelligence. However, it has been widely used in life. For example, in the shopping platform, users can search for a sentence, such as "white dress," and then return the corresponding pictures, text descriptions, video descriptions, and other modal data of the product, which can help users understand the product information; in the locker audio library and video library, you can search for the lyrics or lines to return the name of the song. Or lines to produce the song's title or the frame's position before and after the video appears. However, this problem has not yet been solved in the domestic operating systems field. It still needs the popularization of desktop terminals with domestic operating systems, and users urgently need to query vast amounts of data. Users want to be able to retrieve pictures, audio, video, etc., by typing the words in the localKylin terminal search. openKylin ("Open Kylin")[5], as China's first open-source desktop operating system, has completed the construction of 20+ core components and some eco-applications, marking China's ownership of the operating system components independent of the structure, which is the first time that China has an open-source desktop operating system. This symbolizes that China can independently select operating system components and build the operating system, which fills the gaps at home and abroad. However, we can only search local files, applications, and network resources through Kylin terminal. Some things could be improved in combining the big model with the domestic operating system and retrieving the data stably in the local area. The critical part of the combination is the integration of two different architectures and a suitable database.

2. Materials and Methods

2.1. Overall framework

The whole cross-media retrieval method is composed of ChineseCLIP model[6], SQLite database[7] and Milvus database[8]. The whole system is easy to maintain. During the process of debugging and testing, the AI system responds by calling D-BUS signaling to the model and then to the database. Therefore, the whole method is simple to configure and stable in operation. In this paper, from the perspective of changing the traditional single retrieval result of text matching in openKylin, rich cross-modal retrieval functions such as text search for pictures, text search for audio, text search for video are realized on openKylin, which can satisfy the needs of users in different languages for retrieving pictures, audio, and video, and path search for pictures, which can be able to should be in the collection of huge amount of pictures, audio, and video, without The accuracy of retrieval is not affected. Due to the differences between different modalities, we cannot directly establish the semantic connection between two modalities, so the unimodal retrieval method is inevitably not applicable to the cross-modal retrieval task. In the face of the diverse data forms and rich data contents of the two modalities, how to explore the intrinsic connection of the data has become a difficulty in cross-modal retrieval. Especially in the case of text-based video retrieval[9–12], the video contains rich information forms, such as images, sound, text, etc., which brings great challenges to the characterization and retrieval process of the video.

The traditional method of manual annotation[13,14] not only consumes a lot of time and resources in the process of building the index, but also lacks objectivity and real-time, which drastically reduces the accuracy of retrieval. Therefore, in the case of numerous videos and complex video features, we need to improve the way of processing video information and build an effective and feasible retrieval model to accomplish the task of large-scale video retrieval nowadays. We enter the text to be queried in the search engine, to search for related content of the picture or video is the current cross-modal retrieval task often focuses on the text and picture, to explore the text and picture content matching learning. And this paper explores a comprehensive cross-modal retrieval task - text query based image audio video retrieval task, as shown in Figure 1. The model mainly consists of three stages. In the first stage, the features of picture information, audio information and video information are extracted using free time. In the second stage, the features of the two different modalities obtained from the previous extraction are fed into a vector database respectively, projected into a potential common space and continue processing to obtain a classification result. In the third stage, the user enters a sentence from the searchbox, responds through D-BUS, goes through the text encoder, compares it with the vector database, passes the obtained result ID number to the SQLite database, finds the multimedia address and then gets the final result for return.

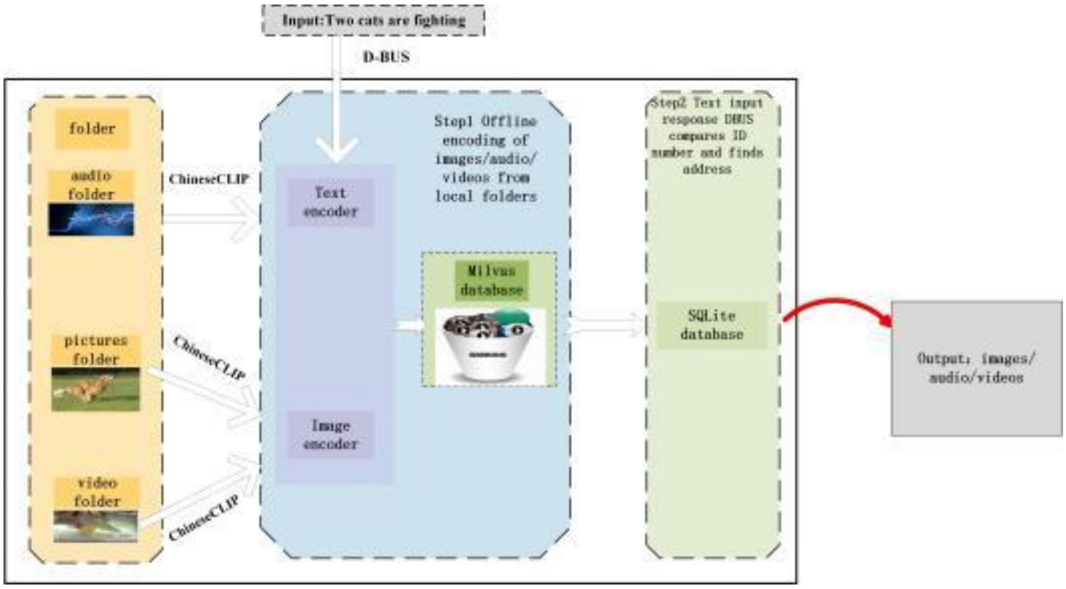

Figure 1. Overall system architecture.

2.2. Text Encoder Module

This research presents a comparative analysis of text processing techniques for severallanguages in the context of user search imagetext. According to Figure 1, the text encoderof ChineseCLIP is employed to encode the present paper. ChineseCLIP, as a pre-training model, is designed for the purpose of language-image comparison during training. This study utilizes the clip-vit-base-patch32[15]model for encoding picture data and employs a customized Transformer model for encoding textual information. The model's performance is optimized by comprehensively utilizing the elements inherent in the pre-existing model.

The characteristics and architecture of the model can be flexibly modified to accommodate specific requirements and scenarios, hence achieving customization. Its purpose is to enhance the existing translation-based API of English and the Stable Diffusion model utilized by most teams in development. The objective is to address the challenges of cultural disparities between Chinese and English. During the training phase, the visual encoder experienced a freeze, while only the Chinese language model underwent fine-tuning. A total of 24 training epochs were conducted on a dataset consisting of approximately 100 million Chinese levels, encompassing approximately 3 billion Chinese characters. The training process involved conducting 24 epochs on a dataset consisting of 100 million Chinese text samples and around 3 billion Chinese graphic data samples. This resulted in the development of a Chinese representative language model incorporating picture information. The ChineseCLIP model and Bert for sequence classification are employed in this study. The Chinese input is tokenized into individual words, and the Bert for sequence classification model is utilized to encode these tokens and generate a textual vector as output. Subsequently, the vector above is employed to search within the embedded vector database to identify the vectors that exhibit the highest degree of similarity. The outcomes of this search are subsequently provided as the results.

2.3. Image Encoder Module

The system has already processed the image in its spare time, as depicted in Figure 1, the user is required to store the image in a vector database. To achieve this idea, the initial step involves utilizing a deep learning model, such as ChineseCLIP, to conduct feature extraction on the image. The primary objective of feature extraction is to convert a picture into a feature vector that contains both semantic and spatial information. This transforma- tion is performed to enhance the efficiency and effectiveness of subsequent processing and management tasks. The image undergoes feature extraction using the CLIPProcessor[1,2]. It is then preprocessed and translated into tensor format by a preprocessor. The resulting tensor is subsequently fed into the ChineseCLIP. When doing image

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 November 2023 doi:10.20944/preprints202311.0185.v1

processing, as depicted in Figure 1, the user must store the image in a vector database. To achieve this idea, the initial step involves utilizing a deep learning model, such as ChineseCLIP, to conduct feature extraction on the image. The primary objective of feature extraction is to convert a picture into a feature vector containing semantic and spatial information. This transformation enhances the efficiency and effectiveness of subsequent processing and management tasks. The image undergoes feature extraction using the CLIPProcessor. It is thenpreprocessed and translated into tensor format by a preprocessor. The resulting tensor is subsequently fed into the ChineseCLIP model for forward propagation. The Chinese-CLIP model is a deep learning model based on the Transformer architecture, known for its robust ability to characterize both images and text. A feature vector of dimensions (1,512) is extracted to obtain the feature vector of an image. This feature vector encompasses both the semantic and spatial information of the image. Semantic information refers to representing objects, situations, or concepts depicted in a picture. On the other hand, spatial information pertains to the depiction of the position, scale, orientation, and other related attributes inside the image. One of the benefits of storing an image as a vector is the ability to employ mathematical operations and similarity metrics to quantify the similarity or dissimilarity between various images. Subsequently, these feature vectors can be keptwithin a Collection in the vector database. A collection refers to a cohesive entity within the Milvus vector database designed to store vector data in a structured manner. The Milvus database is a software system that stores and retrieves high-dimensional vectors efficiently. The Milvus Database offers effective vector indexing and retrieval capabilities to facilitate picture retrieval activities on a wide scale.

2.4. Modal Match Module

Based on the information provided in Figure 2, it is evident that the feature vectors of English text, Chinese text, and images are stored within the Collection object of Milvus. The collection object of Milvus is a fundamental component that plays a crucial role in management and organization of data within the Milvus system. The search operation is executed using the search method of the Collection object in Milvus. This method requires the query vector, search parameters, and a specified limit for the number of returned results. In the process of searching, the L2 distance is employed as a metric of similarity in order to identify the vector that is most similar to the query vector. The user utilizes their local folder to store a collection of images that require retrieval. During their available free time, they opt to store pertinent information about each image, including its name, path, and size, in an SQLite database. This information is then inserted into the IMAGES table of the SQLite database. The IMAGES table consists of three fields: ID, NAME, and PATH. Notably, the ID field serves as the primary key with auto-growth functionality. In the process of insertion, it is not necessary for the user to manually specify the ID value, as the database system will autonomously produce a unique ID value for each newly inserted record. Consequently, the unique ID value will be transmitted to the Milvus vector database. Subsequently, the Milvus vector database will perform feature vector matching between the text and image and subsequently return the unique ID value to the SQLite database. This process enables the retrieval of basic information pertaining to the photo, which is then presented as the output on the user's searchbox.

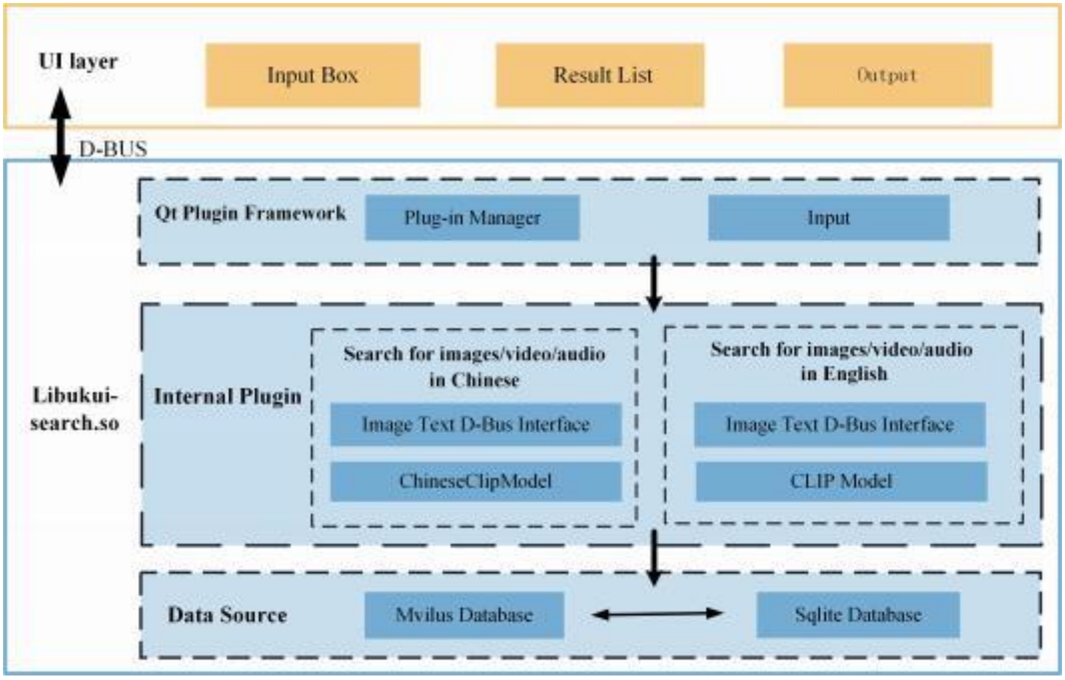

Figure 2. Module Framework.

2.4.1. Libukui-search module

A popular search tool made specifically for the ukui desktop environment is called Libukui-search. Version 3.1-xxx is the most recent version of this program. Aggregated search capabilities for several local resources, including files, text, apps, settings, and notes, are provided via the global search application. A file indexing feature allows the user to have a quick and accurate search experience. Libukui-search is more than just a worldwide search engine. Within the Ukui desktop environment and its related development interface, there are tinier types of file searches available through a local search service. Moreover, the Ukui desktop environment's search function provides a selection of plug-in interfaces based on the Qt plug-in framework. Users can adopt these interfaces to help with the development of the search feature. The approach used in this investigation is file search.

File search is the main methodological strategy used in this investigation. There are two primary categories of search functionality: text content search and file name (folder name) search. Direct search and indexed search are two different types of file searches that are available. Without the capacity to search for text content, a direct search entails iterating over a list of keywords that match. In order to generate a database for direct database searching, indexing search entails navigating the filesystem. Search results maybe obtained in milliseconds with this method. The search results will likely be incomplete or inaccurate while creating an index. The current method is included as a textual search interface packaged inside an external plug-in. It makes use of indexed search to improve word retrieval and to search for desired images based on certain terms.

2.4.2. D-BUS Interaction

To achieve cross-modal graphic retrieval capabilities, the D-BUS interface, the Chinese-CLIP model, and the search capabilities of the openKlylin operating system are combined in the cross-modal graphic retrieval strategy. The cross-modal graphic retrieval approach realizes cross-modal graphic retrieval capacity by combining the D-BUS interface, the CLIP model, and the openKlylin operating system search function. Through the implementation of D-BUS as an inter-process communication mechanism, effective data interchange and cooperation between the CLIP model and the search module are realized. The primary goal is to set up and operate Kylin OS's D-BUS service[16,17], which serves as a conduit for communication between various components. The D-

BUS service enables the search module to communicate in real-time with other modules, such as the ChineseCLIP model for cross-modal matching of words and pictures. The Kylin OS's search module implemented the D-BUS interface, so other components could submit query requests. The ChineseCLIP model is in charge of creating a semantic representation of the picture and connecting textual inquiries with the content of images. The search module computes similarity by receiving the feature vectors supplied by the ChineseCLIP model over the D-Bus interface. The search module organizes the picture results and presents the user with the most relevant photos based on the results of the similarity calculation. It understands that a text query can enable accurate picture and text retrieval for the user. Meanwhile, a loosely linked connection between the search module and the ChineseCLIP model is accomplished using the D-BUS interface, improving the system'sscalability and flexibility.

With this integration strategy, Kylin OS may benefit from a sophisticated cross-modal graphic retrieval solution. It offers a practical way to combine several cross-modal retrieval efforts with a home operating system. A successful resolution.

3. Results

3.1. Experiment environment

This article uses a machine running OpenKylin version 1.0 as the experimental platform. Python and C++are programming languages used. The academic community evaluates graphic retrieval models using publicly available datasets such as the MSCOCO dataset[18–20] to evaluate their accuracy and the EPIC-KITCHENS-100 dataset[21–23], which includes audio and video information. MSCOCO dataset. It has a large amount of image and symbol data to help complete various visual tasks, including detection, segmentation, and image recognition. The dataset includes 400000 fully annotated test photos, 5000 validation images, and 118287 training images. The important features such as category, position, and size of the items in the image are included in the annotation information.

EPIC-KITCHENS-100 dataset, containing audio and visual information; Collecting data involves 4 cities and 45 kitchens; The total duration of the video exceeds 100 hours (full HD, 60fps), with a total frame count of over 20M,including over 90000 action clips, 97 verb categories, and 300 noun categories.

3.2. Classification Results and Analysis

In this paper, we mimic personalized photos, sounds, and videos in user computer folders using the MSCOCO dataset with EPIC-KITCHENS-100. They are used to confirm the code's textual integrity. ChineseCLIP, a preprocessor, was utilized for the search. Store the encoded features in the vector database Milvus, encrypt the multimedia material inside the folder, and then watch for its recovery. As a result, by obtaining the information that fits these two datasets, users' text searches in the Kylin searchbox will be scored based on how quickly and accurately they perform.

3.2.1. Search for images through text

As shown in Figure 3. User can enter "A small squirrel with an umbrella" in English or Chinese by the user in the search field, which is connected to Qt and called over D-BUS. The ChineseCLIP text encoder first loads the backend, then encodes the text, compares it to the vector database, and then outputs the image.

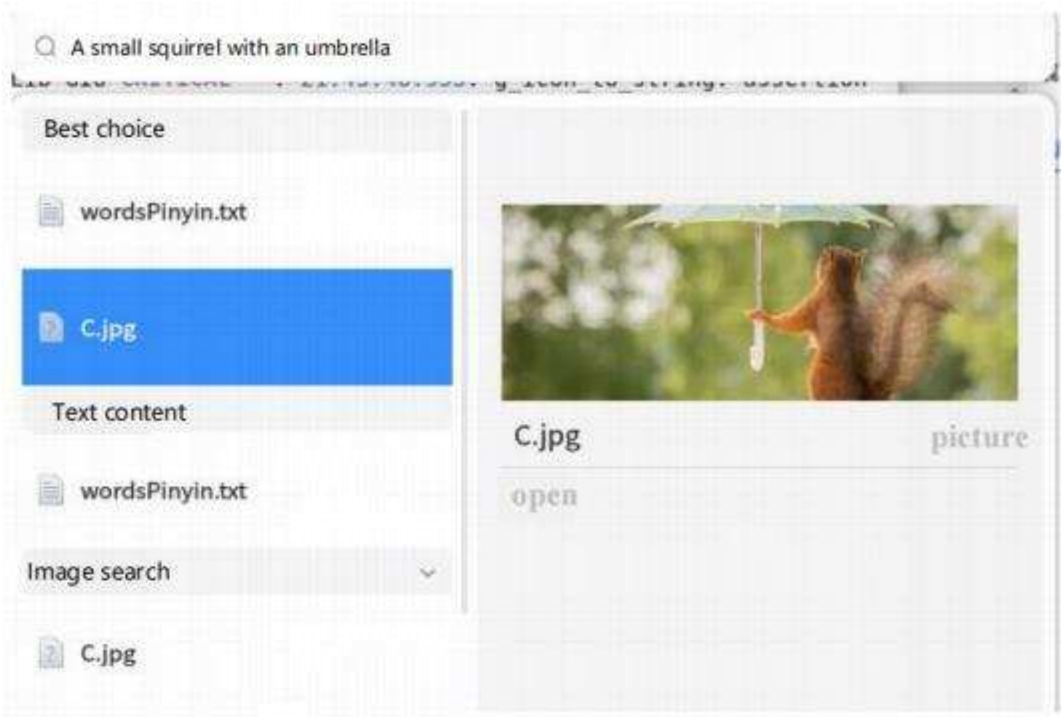

Figure 3. A small squirrel with an umbrella.

3.2.2. Search for audio through text

Modern deep learning models cannot directly analyze audio input in unprocessed forms like ACC, WAV, and MP3. In this work, we will initially use the speech recognition module in Python to capture audio and extract its essential features. The benefit of adopting this is that the voice acquisition will immediately terminate when the user finishes speaking. Import speech_recognition first, then instantiate recognize (). Open the Microphone using Microphone (), adjust the sampling rate, record using the listen function, obtain the audio byte stream data using the get_wav_data() method, and then write the audio in wav format using the write function. The method get_wav_data() obtains the recorded audio byte stream data and uses the write function to output it in WAV format to a file. Following the user's audio data input, this paper uses the text encoder model provided by ChinesCLIP for encoding. It also uses a speech recognition library with Python, speech_recognition[24], and its recognize_sphinx() function to realize essential speech recognition, which is then performed in the speech-to-text, encoder, and vector database. The vector database has the encoder preserved. As seen in the Figure 4, the user input "grass" was filled in by contacting the QT connection and D-BUS response, which finally matched the wav format audio.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 November 2023 doi:10.20944/preprints202311.0185.v1

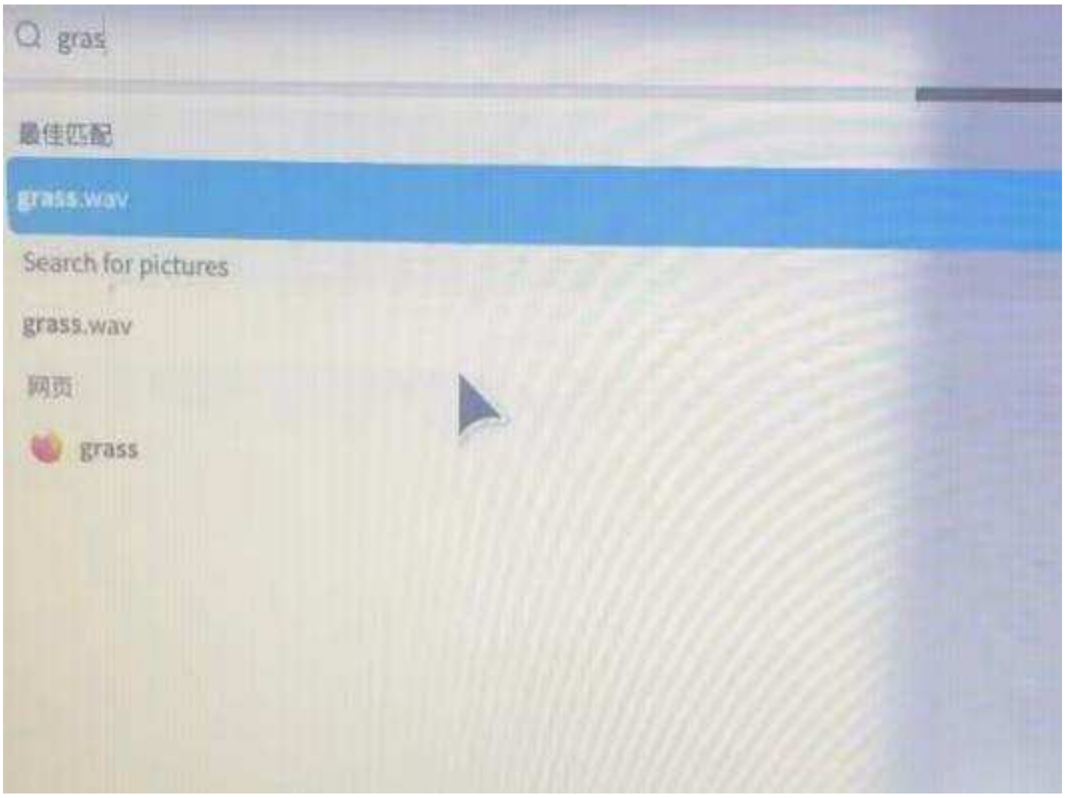

Figure 4. grass.

3.2.3. Search for videos through text

Textual non-video material is used as query input for cross-modal videoCLIP retrieval[25–27]. Even if the feature dimensions of different modalities are identical, it is not feasible to directly compare the similarity of various modal features since other media expressions take distinct shapes. This is why other media data are referred to as different modalities. As a result, joint embedding learning[28] across modalities—a technique used in cross-modal video CLIP retrieval focuses on mapping the feature values of various modalities into a shared space where the similarity connection between different modalities is learned.

The cross-modal video feature fragment retrieval based on natural language text is the main topic of this article. Utilizing a single-sentence query from this paper as the query input. As is shown in Figure 5 for example, "A plane is flying in the sky" video fragment retrieval based on the paper's text falls under the cross-modal video[29] fragment retrieval category.D-BUS calls this query, which is matched with the back-end vector database Milvus and returned by the SQLite database, to precisely locate the descriptive-semantic corresponding segments in the detected video in the Before and After position.

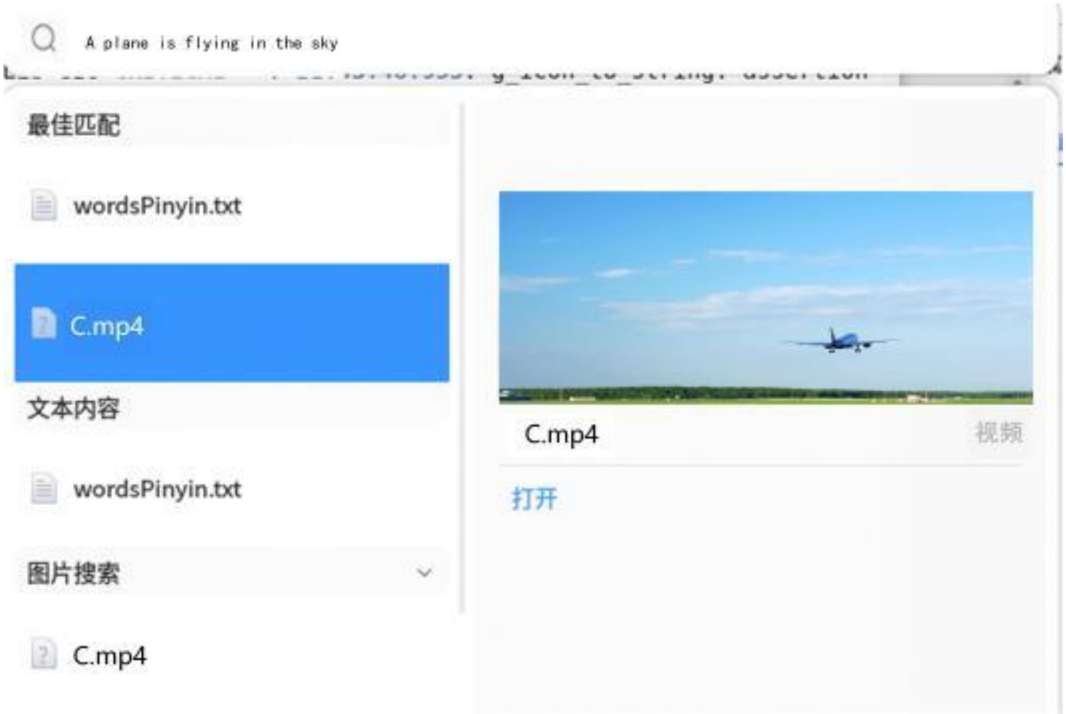

Figure 5. A plane is flying in the sky.

4. Conclusions

To investigate cross-modal retrieval used on the Kylin domestic operating system, we provide an openKylin based cross-modal retrieval approach in this paper. Three distinct search modes are available: text search for images, text search for audio, and text search for video. Enter significant words into the localKylin terminal search to access these modes.

Because we account for the specificity of the local retrieval and record the feature values based on the vector database for comparison and verification, the findings demonstrate that the cross-modal retrieval can reliably recover images, audio, and video from the user's computer. The study's findings demonstrate that the big model retrieval offers a sound concept and a path for integrating the artificial intelligence area with the home operating system. These three recovered sub-directions differ in how much they have improved, and the present modification consists of several user-facing multimedia packages. This study includes cross-modal fundamental environment building, user retrieval phrases, and data consistency inside their folders. The method's hardware and software performance have been evaluated to ensure the system's strong viability. Following Actions: The next stage will be to concentrate on enhancing the retrieval efficiency and training lightweight models to carry out time and memory optimization for openKylin's cross-modal graph retrieval approach.

Author Contributions: Conceptualization, J.Z. and X.X.; methodology, X.L.; software, P.Z., Y.L, and C.Z.; validation, J.Y., L.P. and W.W.; formal analysis, X.X.; investigation, X.X.; resources, J.Z., X.X. and X.L.; data curation, X.X.and X.L.; writing—original draft preparation, X.X.and X.L.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X. and J.Z.; funding acquisition,J.Z.,J.Y. and W.W. All authors have read and agreed to the published version of the manuscript.

Funding: Please add: "This research was funded by Postgraduate Scientific Research Innovation Project of Hunan Province(QL20230220), Haihe Laboratory of Information Technology Application Innovation(22HHXCJC00009),the National Natural Science Foundation of China(61972055),the Research Foundation of Education Bureau of Hunan Province,China(20C0030),the Natural Science Foundation of Hunan Province(2021JJ30734).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jhingan, G.D.; Manich, M.; Olivo-Marin,J.C.; Guillen, N. Live Cells Imaging and Comparative Phosphoproteomics Uncover Proteins from the Mechanobiome in Entamoeba histolytica. International Journal of Molecular Sciences **2023**, 24, 8726.
2. Dehouche, N. Implicit stereotypes in pre-trained classifiers. IEEE Access **2021**, 9, 167936–167947.
3. Qin,J.; Fei, L.; Zhang, Z.; Wen, J.; Xu, Y.; Zhang, D. Joint specifics and consistency hash learning for large-scale cross-modal retrieval. IEEE Transactions on Image Processing **2022**, 31, 5343–5358.
4. Chen, D.; Wang, M.; Chen, H.; Wu, L.; Qin, J.; Peng, W. Cross-modal retrieval with heterogeneous graph embedding. In Proceedings of the Proceedings of the 30th ACM International Conference on Multimedia, 2022,pp. 3291–3300
5. Wang, W.; Liu, X.; Yu, J.; Li,J.; Mao, Z.; Li, Z.; Ding, C.; Zhang, C. The Design and Building of openKylin on RISC-V Architecture. In Proceedings of the 2022 15th International Conference on Advanced Computer Theory and Engineering (ICACTE). IEEE, 2022, pp. 88–91.
6. Yang, A.; Pan,J.; Lin,J.; Men, R.; Zhang, Y.; Zhou,J.; Zhou, C. Chinese clip: Contrastive vision-language pretraining in chinese. arXiv preprint arXiv:2211.01335 **2022**
7. Pawlaszczyk, D. SQLite. In Mobile Forensics–The File Format Handbook: Common File Formats and File Systems Used in Mobile Devices; Springer, 2022; pp. 129–155.
8. Literák, I.; Raab, R.; Škrábal, J.; Vyhnal, S.; Dostál, M.; Matušík, H.; Mako ˇn, K.; Maderiˇc, B.; Spakovszky, P. Dispersal and philopatry in Central European red kites Milvus milvus. Journal of Ornithology **2022**, 163, 469–479.
9. Galanopoulos, D.; Mezaris, V. Are all combinations equal? Combining textual and visual features with multiple space learning for text-based video retrieval. In Proceedings of the European Conference on Computer Vision. Springer, 2022,pp. 627–643.
10. ADLY, A.S.; HEGAZY, I.; ELARIF, T.; Abdelwahab, M. Development of an Effective Bootleg Videos Retrieval System as a Part of Content-Based Video Search Engine. Int. J. Comput **2022**, 21, 214–227.
11. Wu, W.; Zhao, Y.; Li, Z.; Li, J.; Zhou, H.; Shou, M.Z.; Bai, X. A Large Cross-Modal Video Retrieval Dataset with Reading Comprehension. arXiv preprint arXiv:2305.03347 **2023**.
12. Duarte, A.; Albanie, S.; Giró-iNieto, X.; Varol, G. Sign language video retrieval with free-form textual queries. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022,pp. 14094–14104
13. Graves, R.L.; Perrone, J.; Al-Garadi, M.A.; Yang, Y.C.; Love,J.S.; O'Connor, K.; Gonzalez-Hernandez, G.; Sarker, A. Thematic analysis of reddit content about buprenorphine-naloxone using manual annotation and natural language processing techniques. Journal of Addiction Medicine **2022**.
14. Zhang, G.; Sun, B.; Chen, Z.; Gao, Y.; Zhang, Z.; Li, K.; Yang, W. Diabetic retinopathy grading by deep graph correlation network on retinal images without manual annotations. Frontiers in Medicine **2022**, 9, 872214.
15. Dong, X.; Bao,J.; Zhang, T.; Chen, D.; Gu, S.; Zhang, W.; Yuan, L.; Chen, D.; Wen, F.; Yu, N. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-band vit-l on imagenet. arXiv preprint arXiv:2212.06138 **2022**.
16. Gao, Y.; Yu, X. Design and Implementation of Trusted Plug-in Based on Kylin Operating System Platform. In Proceedings of the Journal of Physics: Conference Series. IOP Publishing, 2020, Vol. 1544,p. 012042
17. Chen, Y.; Ma, M.; Yu, Q.; Du, Z.; Ding, W. Road Bump Outlier Detection of Moving Videos Based on Domestic Kylin Operating System. In Proceedings of the Proceedings of the 6th International Conference on High Performance Compilation, Computing and Communications, 2022,pp. 137–143.
18. Bayet, T.; Denis, C.; Bah, A.; Zucker, J.D. Distribution Shift nested in Web Scraping: Adapting MS COCO for Inclusive Data. In Proceedings of the ICML Workshop on Principles of Distribution Shift 2022, 2022
19. Bayet, T.; Denis, C.; Bah, A.; Zucker, J.D. Distribution Shift nested in Web Scraping: Adapting MS COCO for Inclusive Data. In Proceedings of the ICML Workshop on Principles of Distribution Shift 2022, 2022.
20. Chun, S.; Kim, W.; Park, S.; Chang, M.; Oh, S.J. ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO—Supplementary Materials—
21. Damen, D.; Doughty, H.; Farinella, G.M.; Furnari, A.; Kazakos, E.; Ma,J.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision **2022**, pp. 1–23.
22. Lin,N.; Cai,M. EPIC-KITCHENS-100 Unsupervised Domain Adaptation Challenge for Action Recognition 2022: Team HNU-FPV Technical Report. arXiv preprint arXiv:2207.03095 **2022**

23. Lin,N.; Cai,M. EPIC-KITCHENS-100 Unsupervised Domain Adaptation Challenge for Action Recognition 2022: Team HNU-FPV Technical Report. arXiv preprint arXiv:2207.03095 **2022**.
24. Totare, M.R.; Bembade, S.; Chavan, S.; Dighe, S.; Gajbhiye, P.; Thakur, A. SPEECH TO SPEECH TRANSLATION USING MACHINE LEARNING
25. Luo, H.; Ji, L.; Zhong, M.; Chen,Y.; Lei,W.; Duan, N.; Li, T. Clip4clip: An empirical study of clip for end to end videoclip retrieval and captioning. Neurocomputing **2022**, 508, 293–304.
26. Pei, R.; Liu,J.; Li, W.; Shao, B.; Xu, S.; Dai, P.; Lu,J.; Yan, Y. CLIPPING: Distilling CLIP-Based Models with a Student Base for Video-Language Retrieval. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023,pp. 18983–18992
27. Pei, R.; Liu,J.; Li, W.; Shao, B.; Xu, S.; Dai, P.; Lu,J.; Yan, Y. CLIPPING: Distilling CLIP-Based Models with a Student Base for Video-Language Retrieval. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023,pp. 18983–18992
28. Li, K.; Zhang, Y.; Li, K.; Li, Y.; Fu, Y. Image-text embedding learning via visual and textual semantic reasoning. IEEE Transactions on Pattern Analysis and Machine Intelligence **2022**, 45, 641–656.
29. Gorti, S.K.; Vouitsis, N.; Ma,J.; Golestan, K.; Volkovs, M.; Garg, A.; Yu, G. X-pool: Cross-modal language-video attention for text-video retrieval. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022,pp. 5006–5015