ArtemisSearch: A Multimodal Search Engine for Efficient Video Log-Life Event Retrieval Using Time-Segmented Queries and Vision Transformer-based Feature Extraction

Hoang-Phuc Nguyen[1,2], Thuy-Nga Ho[1,2], Minh-Dai Tran-Duong[1,2], The-Luan Nguyen[1,2], Duc-Hao Truong[1,2], Nguyen Huu Quyen[1,2], Phan The Duy[1,2][0000−0002−5945−3712], and Van-Hau Pham[1,2][0000−0003−3147−3356]

[1] Information Security Laboratory, University of Information Technology, Ho Chi Minh city, Vietnam
[2] Vietnam National University, Ho Chi Minh city, Vietnam
{22521129,22520926,22520183,23520899,22520407}@gm.uit.edu.vn, {quyennh,duypt,haupv}@uit.edu.vn

Abstract. In this century, search engines have emerged as a crucial component of the technological landscape. Enterprises require a search engine to retrieve specific information within a particular field. However, they face various challenges due to the rapidly increasing volume of data and the need for effective database management to handle diverse data types. Additionally, the search for data is hindered by difficulties in matching queries with key frames or the limitations in understanding query context. In this paper, we introduce ArtemisSearch, a text-based multimodal search engine designed for temporal event retrieval in videos. In the proposed system, an efficient algorithm for Content-Based Image Retrieval (CBIR) using ViT-H/14 and BEiT3 for feature extraction and an open-source vector database, Milvus, our system efficiently retrieves events by leveraging temporal segmentation of queries and matching embeddings for Artificial Intelligence (AI) applications. Additionally, we developed a web application that allows end users to easily create temporally-aware descriptive queries, efficiently explore top results, and view precise video previews at relevant timestamps. The ArtemisSearch method represents a significant advancement in temporal video retrieval, with potential applications across diverse fields, leading to a smoother and more accurate video search experience.

Keywords: CBIR · ViT-H/14 · BEiT3 · Temporal · Event Retrieval · Search Engine.

1 Introduction

The surge in multimedia content, especially online video, has driven the development of AI models for faster, more efficient data querying. Retrieving video

frames based on textual descriptions of specific moments has become a key research area, advancing global information retrieval [1]. As people increasingly wish to revisit specific scenes from the vast amounts of video they consume, the need for advanced, rapid retrieval systems has grown. This demand calls for solutions that not only offer faster query speeds, but are also adaptable across various platforms [2], [3].

A text-based multimodal search engine is a system that enables users to search for multimedia content, such as images, videos, or audio, using text queries. Unlike traditional search engines that rely solely on textual metadata, a text-based multimodal search engine processes a combination of textual information (such as descriptions, captions, or keywords) and other content-based features like visual, audio, or even contextual data [4], [5], [6]. For instance, in a video retrieval scenario, users might input a text query like "a cat playing piano", and the system would search for relevant videos not only by matching textual metadata but also by analyzing the visual and audio content of videos. The engine might identify visual elements (the cat and the piano) and detect piano sounds. This makes the search more powerful and contextually accurate, even for content that may not be fully or accurately labeled with metadata.

According to research on Vision-Language Pre-training (VLP)[7], numerous advanced vision-language models have emerged, narrowing the gap between pre-trained textual and visual modalities. The advent of Transformer models has demonstrated superior capabilities in processing both language and images. Transformers can learn deep representations from data, fully exploiting the connections between features in images and text, thereby achieving high efficiency in combining and synchronizing information from two different data types. In this context, ViT H/14 [8] and BEiT3 [9] emerge as powerful and versatile choices for vision-language tasks. ViT H/14 [8], a variant of the Contrastive Language-Image Pretraining (CLIP) [10] model, effectively balances performance and accessibility. By learning from a large dataset of image-text pairs and optimizing the cosine similarity between text and image embedding vectors, ViT H/14 [8] not only shows excellent suitability for text-based video retrieval tasks but also excels in computational efficiency. With its optimized design, this model can operate smoothly on diverse hardware systems, from low-configuration computers to those without dedicated GPUs, expanding technology accessibility to a broader audience. Meanwhile, BEiT3 [9] augments its power by deeply integrating language and visual modalities. Using a "masked image modeling" mechanism similar to how language models learn representations of masked words, BEiT3 [9] can learn high-semantic image representations, effectively combining image context with text. This makes BEiT3 [9] an optimal solution for problems requiring complex multimodal processing.

Taking inspiration from these studies, our research introduces an advanced system ArtemisSearch leveraging the potentials of the CLIP-ViT-H/14 [8] and BEiT3 [9] models to extract abstract, semantically rich features from videos in the dataset, while ensuring widespread deployment capability. Although highly performance, this powerful model has not yet reached its full potential for accu-

rate information retrieval. To further enhance the model's performance, we propose using EasyOCR for Optical Character Recognition (OCR)-based queries on text in images or videos. This method is particularly valuable for large datasets containing small details written in multiple languages worldwide or specific traditional characters of a particular country. The extracted characters are arranged into word groups, allowing users to effectively leverage specific contextual attributes such as street names, vehicle license plates, or renowned brand labels to identify the context they are searching for. The accuracy of this data is significantly high when the input is a high-resolution image with a well-defined feature matrix, all optimized to operate efficiently on resource-constrained systems. To fully unlock the model's capabilities and enhance search performance, we augment its functionality with Milvus-based search. Milvus, an advanced open-source vector database management platform designed for large-scale similarity search tasks, helps narrow down the search space for feature vectors, especially for smaller or less accessible images in our current configuration. It provides high scalability, fast search performance, and support for multiple index types, allowing our system to efficiently process complex queries on large video datasets. This strategic integration enables our system to provide faster and more reliable video retrieval, ensuring users receive the best results. Specifically, within the context of the LifeLogs Retrieval challenges at the AI Challenge Ho Chi Minh City 2024, our approach demonstrates its effectiveness in accurately resolving all 30 queries of the final round.

Our approach encompasses the entire preparation process for both text-based and video preview queries, addressing the multifaceted nature of multimedia content retrieval. By combining advanced AI models, efficient indexing, and user-friendly interfaces, our system aims to make the vast sea of video content more navigable and meaningful for users seeking to relive their memories or explore visual information. The following sections of this paper will delve into the architecture and core components of the system, as well as the deployment of applications that support user interaction and easy access to large volumes of video, including the retrieval of events within frames. We explore how our integrated approach tackles the challenges posed by the ever-growing volume of video data, making it easier for users to find and revisit the moments that matter most to them.

2 Methodology

This section offers an overview of video processing and the architecture employed in our ArtemisSearch system.

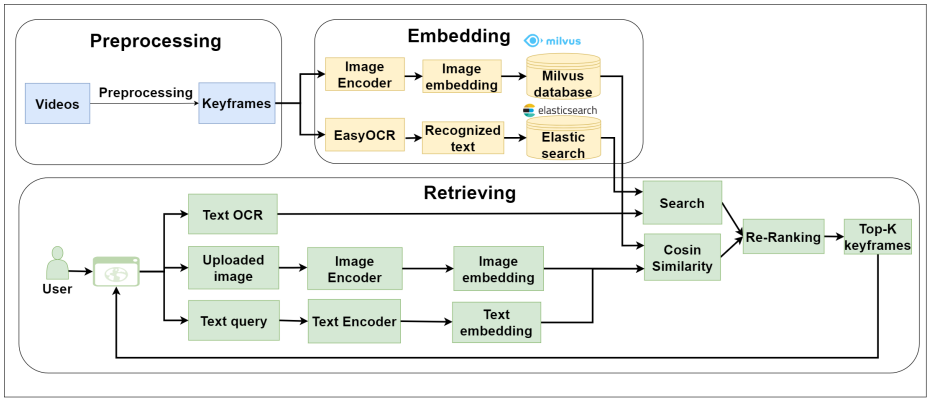


Fig. 1: The Architecture overview of ArtemisSearch System.

2.1 System Architecture Overview

As shown in Fig. 1, our ArtemisSearch system showing three main components: (1) Preprocessing module for extracting keyframes from videos, (2) Embedding module utilizing Milvus for image embedding storage and Elasticsearch for OCR text indexing, and (3) Retrieving module that combines image similarity search with text-based retrieval to produce top-K keyframes based on user queries. The results from these channels are combined using our scoring mechanism to produce the final ranked list of relevant keyframes. This integrated approach ensures comprehensive coverage of both visual and textual content while maintaining efficient retrieval performance.

2.2 Video Preprocessing

FFmpeg is an open-source software framework which provides a suite of tools for handling multimedia data. At the first stage of our system pipeline, known as video preprocessing, we utilize FFmpeg to extract keyframes based on the level of scene changes within the video. This method employs a filter to identify frames that exhibit significant changes compared to previous frames, enabling us to select representative frames that capture these transformations.

After extracting the keyframes, we resolve the issue of duplicate frames by using perceptual hashing to compare the visual characteristics of each keyframe. If the distance between hash values is within a predefined threshold, the keyframes are deemed duplicates and removed. The remaining frames will be processed in the following indexing stages.

2.3 Multimodal Retrieval Models

Our study employs two state-of-the-art models for multimodal retrieval, which are utilized independently to compare their respective performances in video content analysis:

CLIP-ViT-H/14 The CLIP-ViT-H/14 [8] model combines a vision transformer architecture with a contrastive learning framework to enable the model to understand and relate text descriptions to visual content effectively. The model is trained on a diverse dataset containing image-text pairs, allowing it to learn rich representations of visual features. In our approach, we utilize the pre-trained CLIP-ViT-H/14 [8] model to extract semantically rich features from video frames.

To implement this, we follow these steps:

1. **Video Frame Extraction:** We extract frames from videos in our dataset at a predefined interval (e.g., every second) to ensure we capture a representative sample of the video's content.
2. **Feature Extraction:** Each extracted frame is fed into the CLIP-ViT-H/14 [8] model, generating a high-dimensional feature vector that captures the visual semantics of the frame.
3. **Embedding Storage:** The resulting feature vectors are stored in a Milvus database for efficient retrieval during the query phase.

BEiT3 As an alternative approach, we evaluate the BEiT3 (Bidirectional Encoder representation from Image Transformers) model excels in understanding contextual relationships within images.

The implementation process includes:

1. **Feature Encoding:** Processing of video frames through the BEiT3 [9] model to obtain feature vectors. sample of the video's content.
2. **Embedding Storage:** Storage of BEiT3-generated feature vectors in a separate Milvus collection for comparative analysis.

Both models are used independently to extract features from the same set of video frames, allowing for a direct comparison of their effectiveness in our retrieval tasks.

2.4 Score Combination And Re-Ranking Results

Our novel approach to score combination and result re-ranking integrates temporal[11] relevance and textual information to enhance retrieval accuracy:

Temporal-aware Score Combination To combine the scores of retrieval results from multiple query descriptions[11] for each event frame based on specific temporal conditions, we should establish relationships for frames that are close together in a video during the frame extraction process. We propose an advanced re-ranking algorithm that incorporates temporal context[11] into the retrieval process, as shown in Algorithm 1. This algorithm dynamically adjusts scores based on temporal proximity, enhancing the relevance of temporally coherent results.

OCR-Enhanced Score Refinement To further refine our retrieval results, we integrate OCR capabilities using EasyOCR. This addition allows us to incorporate textual information present within video frames into our scoring mechanism:

1. **Text Extraction:**Application of EasyOCR to extract textual content from keyframes.
2. **Score Fusion:**Integration of OCR-derived textual relevance with visual similarity scores using a weighted combination:

$$\text{FinalScore} = 0.7 \times \text{QueryScore} + 0.3 \times \text{OCRScore} \tag{1}$$

Algorithm 1 Enhanced Re-ranking with Temporal Integration[11]

Require: List of retrieval keyframes A, List of query descriptions Q
Ensure: Re-ranked list of keyframes B
Initialize dictionary D for frame scores
Set temporal relevance window T
for each query q_i in Q **do**
  Perform Milvus similarity search for q_i, retrieve top k frames R with scores
  for each frame a in R **do**
    if a in D **then**
      $D[a] = \frac{2}{3} \times (D[a] + \text{new\_score}(a))$
    else
      $D[a] = \text{new\_score}(a)$
    end if
  end for
end for
Initialize list B
for each frame a in D **do**
  if a is within time window T of frames in B **then**
    $D[a] += \delta$ // Temporal relevance boost
  end if
  Apply OCR-based score adjustment (see Section OCR filter)
  Append a to B
end for
Sort B in descending order by score
return top 100 frames from B

3 ArtemisSearch System Overview

This section describes the implementation details of our system's service layer and user interface components.

3.1 Data Service Layer

In our approach, Data Service Layer acts as the core infrastructure component, managing the input and output data processing. We have chosen the FastAPI

framework to optimize the performance and throughput of the API interactions between the system and the user interface within this service layer.

3.2 User Interface

The UI of ArtemisSearch consists of several key components, as shown in Fig. 2:

1. Input area for search requirements
2. OCR Filter section for text-based filtering
3. Results area displaying matched video frames
4. Frame Information section showing detailed frame information and context
5. Video Player component for playback from selected frames

These components work together to form a comprehensive tool for searching and analyzing video content based on textual or image queries.

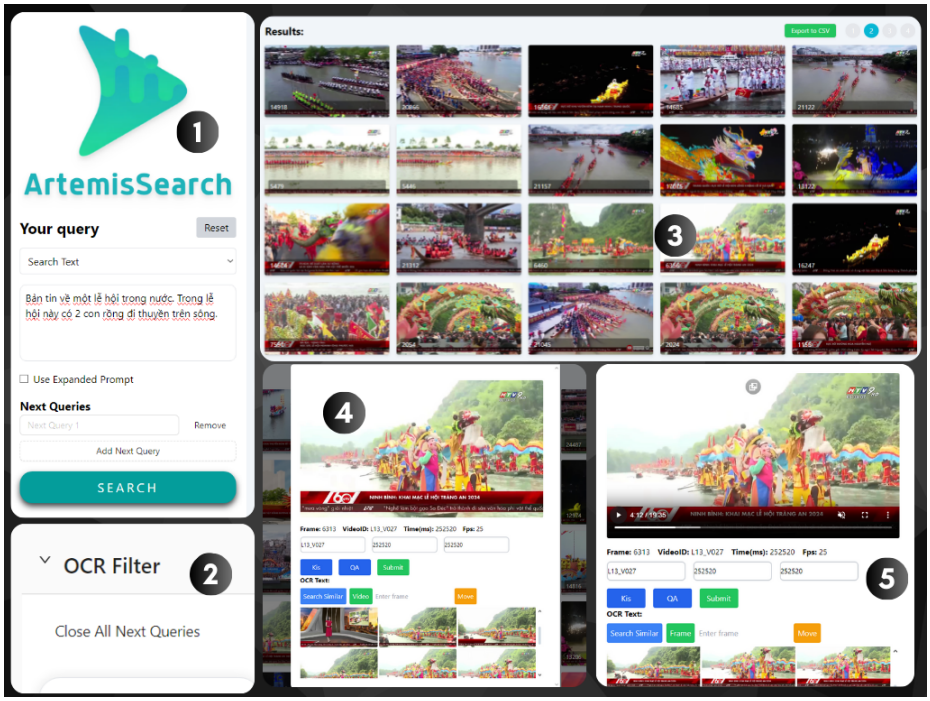


Fig. 2: The User Interface of ArtemisSearch.

3.3 Scenarios of Usage

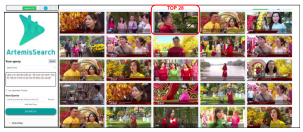
In this section, we showcase examples of how our system retrieves relevant videos from a large collection using text queries from the 2024 Ho Chi Minh City AI

Challenge, which focuses on AI solutions for real-world issues in Ho Chi Minh City. The queries cover topics like culture, tourism, education, health, and the environment.

The 2024 Ho Chi Minh City AI Challenge dataset The dataset used for this year's competition comprises news and events reported across various media channels within the past 18 months. The dataset includes:

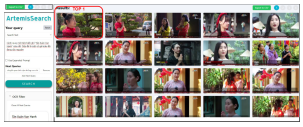
- **Video:** The total duration of video content is 500 hours, divided into three batches. Batch 1 consists of 100 hours, while both Batch 2 and Batch 3 contain 200 hours each.
- **Keyframe:** These frames serve as representative snapshots capturing specific events at particular time points.
- **Metadata:** The dataset also includes descriptive, spatial, and temporal information that corresponds to the videos.

OCR Integration in ArtemisSearch: The integration of OCR into our image search process has led to a substantial improvement in performance. Our comparative analysis reveals a marked enhancement in search accuracy and result ranking when OCR is employed. As demonstrated in Fig. 3(a), without OCR, the target image is ranked at position 28, indicating suboptimal search performance. In Fig. 3(b), Frame 7131 is shown, which contains the OCR-extracted text .This frame appears before the top 28 results at frame 7253 and plays a critical role in the OCR-enhanced search. In contrast, Fig. 3(c) shows that with OCR integration, the system accurately identifies the correct image as the top result. This significant improvement highlights the essential role of OCR in enhancing search relevance and overall image retrieval accuracy.



(a) Without OCR: Target image at rank 28

(b) Frame 7131 contains the OCR text, which is a previous frame of the top 28st - frame 7253.



(c) With OCR: Target image at rank 1

Fig. 3: Comparison of search results with and without OCR integration.

The OCR integration allows the search engine to effectively parse and utilize textual information embedded in images, thereby providing a more comprehensive and context-aware search capability.

Utilizing image-based input with ArtemisSearch: As shown in Fig. 4, video search process using image input with ArtemisSearch: (1) Google search to select an image similar to the desired video content. (2) Using the selected image as input for search on ArtemisSearch. (3) Search results display video frames deemed closest to the descriptive image. This process allows users to search video content based on a reference image, enhancing efficiency in locating specific scenes or moments within videos.

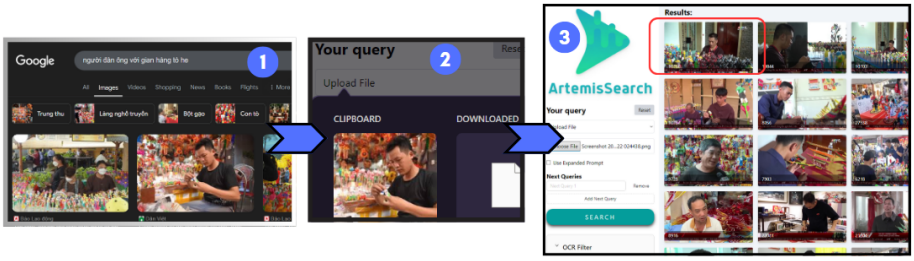
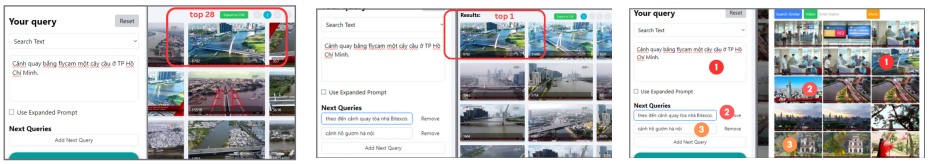


Fig. 4: An example of utilizing image-based input with ArtemisSearch.

Enhanced Temporal Search Relevance through Multiple Queries: Our research demonstrates a significant improvement in video frame search performance and relevance through the implementation of multiple queries. This approach enhances the temporal context of the search from the sudy of Zhang Gengyuan et al. [11], leading to more accurate results. The comparison in Fig. 5 demonstrates that using multiple queries results in better performance compared to a single query.



(a) Single query: Target frame at rank 28

(b) Multiple queries: Target frame at rank 1

(c) Query descriptions for sequential events

Fig. 5: Comparison of search results: Single vs. Multiple queries

4 Conclusion

This paper presents ArtemisSearch, an innovative multimodal video retrieval system that effectively addresses the growing challenges in managing and searching large-scale video content. Our system makes several key contributions to the field of multimedia information retrieval. First, we successfully integrated state-of-the-art vision-language models (CLIP-ViT-H/14 [8] and BEiT3 [9]) with OCR capabilities to create a comprehensive retrieval solution. The combination of these technologies enables our system to understand both visual semantics and textual information present in video frames, significantly improving search accuracy and relevance. Second, our temporal-aware score combination and reranking approach demonstrates the importance of considering temporal relationships[11] between video frames. This novel scoring mechanism, enhanced by OCR-based refinement, helps deliver more contextually relevant search results while maintaining computational efficiency. Third, the system's architecture, built on Milvus for vector similarity search and Elasticsearch for text indexing, proves to be both scalable and efficient. The preprocessing pipeline, utilizing FFmpeg and perceptual hashing, effectively handles the challenge of extracting and managing representative keyframes while eliminating redundancy. Overall, ArtemisSearch represents a significant step forward in making video content more accessible and searchable, particularly beneficial for applications requiring precise moment retrieval in large video collections.

5 Future Work

While our current approach demonstrates promising results, several compelling research directions remain to be explored. One particularly intriguing avenue involves the integration of audio-based search capabilities into our existing framework. As we have observed in our preliminary investigations, many queries inherently contain voice-related information that could potentially enhance both search speed and precision. Building upon the groundbreaking work of Le et al. [12] in audio-based information retrieval, we envision developing a multimodal search system that seamlessly combines textual and audio features. Furthermore, we recognize the potential for enhancing query-information relationships through advanced query reformulation strategies. Drawing inspiration from the innovative approach proposed by Lokoč et al. [13], we plan to implement a context-aware query expansion mechanism. Preliminary experiments suggest that such reformulation strategies could significantly improve search accuracy.

Acknowledgements This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

1. Newton Spolaôr, Huei Diana Lee, Weber Shoity Resende Takaki, Leandro Augusto Ensina, Claudio Saddy Rodrigues Coy, and Feng Chung Wu. A systematic review

on content-based video retrieval. *Engineering Applications of Artificial Intelligence*, 90:103557, 2020.

2. Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu. Deep learning for video-text retrieval: a review. *International Journal of Multimedia Information Retrieval*, 12(1):3, 2023.
3. Meng Liu, Liqiang Nie, Yunxiao Wang, Meng Wang, and Yong Rui. A survey on video moment localization. *ACM Computing Surveys*, 55(9):1–37, 2023.
4. Guanfeng Wu, Abbas Haider, Ivor Spence, and Hui Wang. Multi modal fusion for video retrieval based on clip guide feature alignment. In *Proceedings of 2024 ACM ICMR Workshop on Multimodal Video Retrieval*, pages 45–50, 2024.
5. Tayfun Alpay, Sven Magg, Philipp Broze, and Daniel Speck. Multimodal video retrieval with clip: a user study. *Information Retrieval Journal*, 26(1):6, 2023.
6. Ye Zhu, Yu Wu, Nicu Sebe, and Yan Yan. Vision+ x: A survey on multimodal learning in the light of data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
7. Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.
8. Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
9. Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.
10. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
11. Gengyuan Zhang, Jisen Ren, Jindong Gu, and Volker Tresp. Multi-event video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22113–22123, 2023.
12. Thanh-Thien Le, Linh The Nguyen, and Dat Quoc Nguyen. Phowhisper: Automatic speech recognition for vietnamese. *arXiv preprint arXiv:2406.02555*, 2024.
13. Jakub Lokoč, Zuzana Vopálková, Patrik Dokoupil, and Ladislav Peška. Video search with clip and interactive text query reformulation. In *International Conference on Multimedia Modeling*, pages 628–633. Springer, 2023.