Glimpse: A Multimodal-based Transforming Image Collection with Vector Database

Huu-Tuong HO *, Minh-Tien PHAM *, Quang-Duong TRAN * Quang-Huy PHAM *, Luong Vuong Nguyen*
*Department of Artificial Intelligence, FPT University, Danang 550000, Vietnam
Email: {tuonghhde170471, tienpmde170231, duongtqde160638, huypqde170011}@fpt.edu.vn, vuongnl3@fe.edu.vn

***Abstract*—In the rapidly advancing technological landscape, smartphones have transformed photo and video capture, flooding us with multimedia data. Managing and finding specific media is a big challenge. This study proposed Glimpse, an AI-powered platform that eases media management and retrieval. Glimpse leverages cutting-edge AI technologies to analyze, extract, and store meaningful features from images and videos, ultimately enhancing the performance of the search engine. Specifically, Glimpse includes the generation of embeddings from images, enabling subsequent text-based queries to yield more accurate and contextually relevant results. Furthermore, the system optimizes geolocation data for media, facilitating the swift and accurate retrieval of content associated with specific locations or events. We deploy Glimse with the image datasets collected by our research groups. We then estimate the performance of the proposed model in extracting and retrieving information from images in these datasets. As a result, Glimpse excels at detecting and categorizing images either understanding the query method at all. This demonstrates that Glimse is efficient and intelligent in organizing and retrieving photos and videos based on people.**

***Index Terms*—Multi-modal, Content-based Retrieval, Multimedia Retrieval,**

I. Introduction

The rapid advancement of technology has changed the way we capture, share, and store our memories through photos and videos. With the proliferation of smartphones and the need to document our lives, we have found ourselves flooded with multimedia data. This has led to a new set of challenges in media management and retrieval. As digital footprints increase, it becomes essential to have efficient tools to navigate this vast ocean of multimedia content. To meet this need, we introduce Glimpse, an AI-powered system designed to revolutionize how we manage and retrieve multimedia data. Glimpse combines advanced AI technologies, vector database integration, and content-based image retrieval (CBIR) to offer a comprehensive and multimodal approach to media organization and access.

- Multimodal combination: Glimpse utilizes state-of-the-art AI algorithms and deep learning models with BLIP to understand the query, analyze, extract, and store meaningful features from images and videos, resulting in improved search engine performance.
- Vector database integration: Glimpse's AI system generates embeddings from images, allowing users to perform text-based queries and obtain accurate, contextually relevant results.
- Map-based Queries by geolocation data: Glimpse efficiently organizes and retrieves content based on specific locations or events, making it easier for users to find and revisit their memories.
- Face detection : Glimpse's AI technology can detect and categorize faces within albums, providing a versatile tool for organizing and retrieving photos and videos based on individuals.
- Intuitive and intelligent media management: Glimpse's AI-powered system simplifies the complex process of media management and retrieval, offering a glimpse into the future of efficient and context-aware media organization.

In an age of exponential growth in multimedia data, Glimpse emerges as a pioneering solution to simplify the intricate process of media management and retrieval. By harnessing the full potential of AI, Glimpse provides an intuitive and intelligent means to navigate vast collections of photos and videos, offering a glimpse into the future of efficient and context-aware media organization.

The remainder structure of the manuscript is as follows. In Section II, we provide the related work that describes the studies relevant to image processing, focusing on retrieval and query search techniques. Section III provides a system overview and user interface. Section IV covers the methodologies underlying the transformative capabilities of Glimpse, including Named entity recognition (NER), Vision-language Pre-training, and Face recognition, as well as how Glimpse integrates vector databases and optimizes map-based queries. Finally, Section V showcases the dataset, evaluation metrics, setting, real-world applications of Glimpse, and how it revolutionizes multimedia data management.

II. Related work

The Glimpse online platform, introduced in this study, aims to provide a transformative solution for image collection management with a vector database. To contextualize Glimpse and understand its place in the field of multimedia retrieval, we delve into the related work in this domain, covering key milestones and recent developments.

The QBIC (Query by Image and Video Content) system, presented in the seminal work by Flickner et al. in 1995, marked a significant advancement in multimedia retrieval [1]. QBIC was one of the early attempts to allow users to search for images and videos using visual content as queries. It employed techniques for color, texture, and shape analysis to index and

Corressponding author: vuongnl3@fe.edu.vn

979-8-3503-3094-6/24/$31.00 ©2024 IEEE

retrieve multimedia data. [2] introduces the vitrivr system and explores multimodal approaches to multimedia retrieval. This system combines visual, audio, and textual information to improve the retrieval of multimedia content. Explore the advancements made in vitrivr can offer valuable insights and methodologies that can be applied to Glimpse. In the paper presented at ICCV 2021 [3], Changpinyo et al. explored the use of multimodal queries for image retrieval. This work is particularly relevant to Glimpse, as it highlights the growing importance of multimodal approaches that combine textual and visual information to enhance image retrieval. In this research, the authors proposed a novel framework for querying images by providing both textual descriptions and a reference image, improving the precision of retrieval results. This approach aligns with Glimpse's multimodal capabilities and reflects the ongoing effort to make image and video retrieval more user-friendly and effective.

In "Myscéal", an experimental interactive lifelog retrieval system presented at the Lifelog Search Challenge (LSC'2020), the authors introduced a lifelog retrieval system that helps users explore their personal lifelogs [4]. This work emphasizes the importance of user interactivity and engagement in multimedia retrieval. Similarly, "Memento" was introduced as a prototype lifelog search engine presented at LSC'21, which extends the concept of lifelog retrieval [5]. This system addresses the challenge of managing and retrieving multimedia content from personal lifelogs. Besides, Memento also emphasizes the need for efficient indexing and retrieval mechanisms. It underscores the importance of keeping up with the growing volume of multimedia data generated by users and the need for innovative lifelog management solutions. In [6] presented "LifeInsight" is an interactive lifelog retrieval system. It focuses on providing spatial insights and query assistance to users in managing and retrieving multimedia lifelog data. This research emphasizes the importance of considering the spatial aspect of multimedia content, which aligns with Glimpse's vector database and its ability to manage images based on spatial relationships. The study in [7] introduced "lifeXplore," a lifelog retrieval system that participated in the challenge. This challenge allowed different systems to compete and showcase their capabilities in retrieving personal lifelogs. While Glimpse is primarily designed for transforming image collections, exploring the lifelog retrieval landscape and the systems that participate in such challenges can offer valuable insights into improving the retrieval and management of multimedia collections.

In summary, Glimpse, with its emphasis on transforming image collections using a vector database and its multimodal capabilities, draws inspiration from and builds upon these key milestones and developments to offer an innovative solution for image collection management.

III. Overview of Glimpse

A. System Overview

Figure 1 provides a visual representation of the system's operation. This system takes a user's text query, which can include information like a person's name, time, and location. Upon receiving the query, Glimpse dissects it into two key components: the query's semantic content and additional attributes for future filtering, such as time, location, and human names. Each image in the photo albums is embedded and stored within the Qdrant vector database, with its metadata also saved in the payload of the vector point to enable subsequent filtering. Furthermore, Glimpse employs Face detection and recognition modules, which store individual identities in a separate vector database. The semantic aspect of the query undergoes text encoding and is then fed into the Qdrant Database for similarity searches and ranking, complemented by geolocation, human name, and time filtering. The system then returns the refined results to the user.

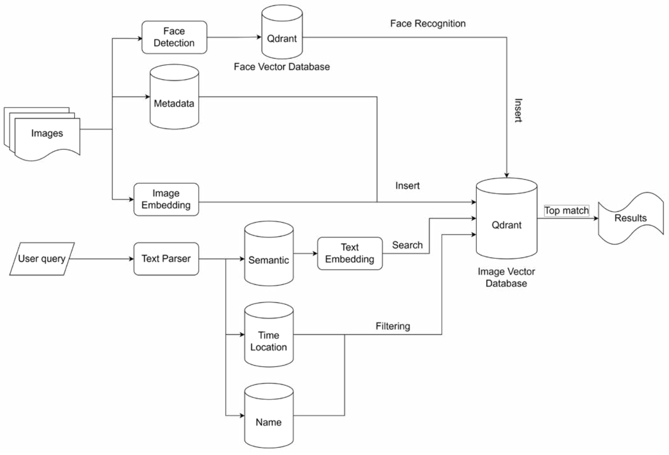
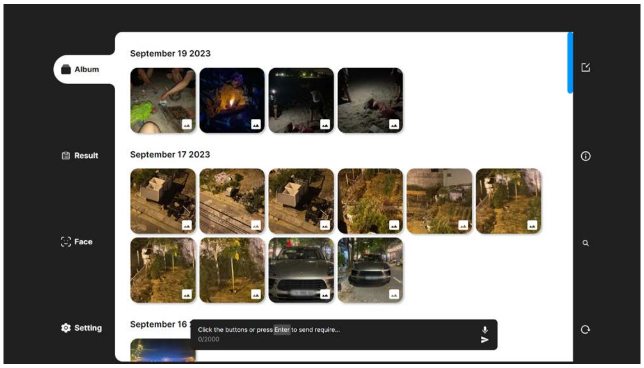


Fig. 1. The architecture of online platform Glimpse

B. User Interface



Click the buttons or press Enter to send require...
0/2000

Fig. 2. Screenshot of user interface of Glimpse online Platform

Displayed in Figure 2 is the overview user interface of our Glimpse system. The front end of our system was constructed utilizing Next.js[1], a versatile React framework that provides a toolkit for crafting high-performance web applications. Positioned on the left-hand side of the user interface is a sidebar housing navigation elements, facilitating user exploration of

[1] https://nextjs.org

photos based on timelines, search results, facial recognition, and advanced settings. The central-bottom area is dedicated to user input, allowing them to enter their queries. The search results are subsequently presented in the results tab for user perusal.

IV. Methodology

A. Named entity recognition (NER)

In information extraction, named entity recognition—also known as entity chunking or extraction is a widely used technique for identifying and segmenting named entities as well as classifying or categorizing them under different specified classifications such as organizations, persons, dates, and so on [8]. Table I lists some commonly used types of Named Entities.

TABLE I
Typical categories of named entities in common usage [9]

NE categories	Examples
ORGANIZATION	Google, Microsoft
PERSON	Elon Musk, Cristiano Ronaldo
LOCATION	Ho Chi Minh City, Da Nang City
DATE	April, 2023-09-30
GPE	South East Asia, Midlothian

Recent advancements in Natural Language Processing (NLP), particularly the introduction of pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers), have revolutionized the field, pushing the boundaries of NER performance. In the Glimpse system, we explore the potential of fine-tuning BERT, a state-of-the-art pre-trained language model, for Named Entity Recognition tasks. BERT has demonstrated remarkable capabilities in capturing contextual information from text [10], and achieving state-of-the-art results in a variety of NLP activities has proven to be incredibly effective when fine-tuning it for specific tasks. By harnessing the power of BERT for NER, we aim to address some inherent challenges in traditional NER systems, such as handling rare entities, coping with context-dependent entity recognition, and achieving high accuracy on complex and diverse textual data.

B. Semantic Search

Within our system, we implement Bootstrapping Language-Image Pre-training (BLIP) technique. This pioneering framework falls under the Vision-Language Pre-training (VLP) category and is noted for its advanced proficiency in both comprehending and generating tasks that intertwine vision and language, as demonstrated by its leading performance in the zero-shot image-text retrieval task using the Flickr30k dataset. [11].

The method for semantic searching within BLIP mirrors the modus operandi of the Contrastive Language-Image Pre-training (CLIP) model. The fundamental principle driving both frameworks is to ensure that images and text queries ”speak the same language” by transforming them into corresponding vectors within a singular vector space [12]. This conversion allows for the comparison of similarities by translating the images' contextual data — which binds visual and linguistic elements and the text queries — which incorporate the relationship of essential terms—into vectors of the same scale. Subsequently, Qdrant [2], a vector database optimized for similarity search, performs the similarity computation, delivering a ranked list of images that are most relevant to a given query description. To achieve this, cosine similarity is utilized as the distance metric for comparing embeddings in the vector space as equation 1.

$$\text{cos\_sim}(A,B) = \frac{A \cdot B}{||A|| * ||B||} = \frac{\sum_i^n A_i B_i}{\sqrt{\sum_i^n A_i^2}\sqrt{\sum_i^n B_i^2}} \tag{1}$$

It is important to note that our system normalizes the embeddings before storing them in the database and configures the database to employ the Dot product operation as shown in equation 2. This configuration ensures the use of cosine similarity with Dot product, a highly efficient operation, thanks to Single Instruction, Multiple Data (SIMD) technology, further enhancing the system's performance.

$$\text{dot\_product}(A,B) = A \cdot B = \sum_{i=0}^{n-1} A_i B_i \tag{2}$$

Where:

- A is 1st non-zero vector
- B is 2nd non-zero vector
- n is dimension of the vector space
- a_i is component of vector A
- b_i is component of vector B

C. Face recognition

Face recognition is a machine technology that uses human facial features to determine a person's identity. This technology has developed rapidly in recent years and is increasingly widely used in many sectors. Facial recognition has been researched since the 1960s. Early facial recognition methods relied on geometric features, such as the distance between the eyes, nose, and mouth.

With Glimpse, we use two popular and very powerful algorithms Multi-task Cascaded Convolutional Neural Networks (MTCNN) for detected faces and FaceNet for Face Embeddings. MTCNN is a face detection model developed by Zhang et al. (2016), It's a machine learning-based model that can detect faces in photos and videos with high accuracy [13]. MTCNN uses a simple neural network to detect regions likely to contain faces in the image, next it uses a more complex neuron network to refine the position of the previous faces. Finally, MTCNN uses a neural network to reformat faces into a certain size and ratio [13].

In Facenet, face embedding has generated by a Convolutional Neural Networksk (CNN). CNN in Facenet learn how to extract features of faces [14]. This features use for create a vector with 512 dimensions [3].

[2] https://qdrant.tech

[3] https://pypi.org/facenet

D. Geo-location filtering with density-based clustering algorithm

At Glimpse, our approach involves employing machine learning cluster algorithms to organize and visually represent image locations on a map based on their respective longitude and latitude coordinates. Following extensive research into geospatial data handling, we have determined that the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is the most suitable option for our needs. [15]

DBSCAN is classified as a density-based clustering algorithm, which means it groups data points together based on the density of neighboring data points [16]. This characteristic makes it an ideal choice for clustering geo-location data, as it can effectively identify clusters of points that are situated closely in space, regardless of their size or shape variations. Furthermore, DBSCAN exhibits a notable resilience to noisy data, as it can effectively filter out outliers that do not belong to any dense cluster [16]. Figure 3 provides an illustration of how DBSCAN clusters data points.

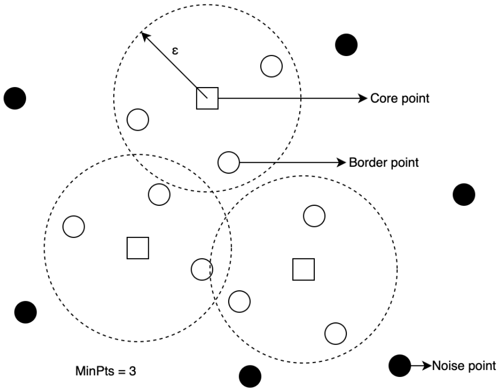


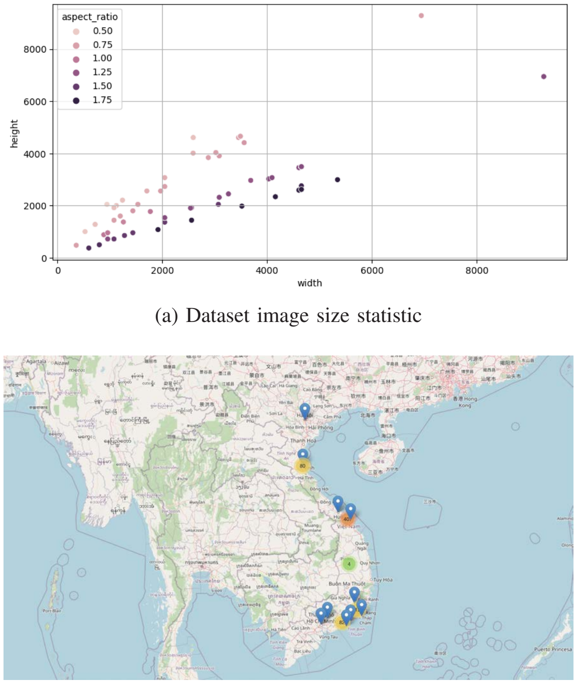
Fig. 3. An example of DBSCAN in action

V. Experiments

A. Dataset

Our photo albums dataset was assembled through a crowdsourcing effort, aimed at curating a diverse collection of images. In total, the dataset comprises 700 images, each of which is accompanied by essential metadata. Notably, this metadata includes geospatial information in the form of a two-dimensional vector, consisting of longitude and latitude coordinates. The majority of the images within our dataset were captured using smartphone devices, reflecting the ubiquitous nature of these devices in modern photography. This prevalence underscores the relevance of our dataset in the context of contemporary image analysis and computational photography.

In Figure 4(a), we present a statistical analysis of the image resolution and aspect ratios of all the photos within our dataset. This analysis reveals insights into the distribution of image sizes, which can be instrumental in various image-processing tasks. Notably, we encountered two "outlier" images within our dataset, characterized by huge file sizes. These images were captured using the super-resolution mode available on select Android smartphones. The presence of such outliers exemplifies the diversity in the image capture process and provides an interesting avenue for further analysis and research.



(a) Dataset image size statistic

(b) Map visualization

Fig. 4. The details describe of collected image dataset

Figure 4(b) illustrates the geographical distribution of the collected data points within our dataset. It is evident that the dataset is primarily distributed across the region of Vietnam. The geospatial aspect of our dataset is not only valuable for location-based analysis but also adds an interesting dimension for understanding the cultural and environmental context in which these images were captured.

B. Experiments settings

To have a better look at the data's apportion and search the images or videos belonging to a certain location, we cluster the geolocation data using the DBSCAN algorithm, which requires two parameters: Min Points (MinPts) and Epsilon (Eps). We set MinPts to 4 for 2D data and determine Eps using the Elbow method [15] [17], which gives us a value of 0.15 for our dataset.

Once we have found the values for MinPts and Eps, we cluster the data and show the clustered data points in Figure 4. To display the map, we use the Geolocator library to convert location names into coordinates. We then calculate the haversine distance 3 between the location's coordinate and the centroids' coordinate and return the points that belong to the cluster of the nearest centroid. Haversine distance formula:

$$(\mathrm{d}) = 2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\mathrm{lat}}{2}\right) + \cos(\mathrm{lat}_1)\cdot\cos(\mathrm{lat}_2)\cdot\sin^2\left(\frac{\Delta\mathrm{lon}}{2}\right)}\right) \quad (3)$$

where:

- d - The Haversine distance between two points (in kilometers, miles, etc., depending on the radius used).
- r - The radius of the Earth (in kilometers, miles, etc.).
- $\Delta\mathrm{lat}$ - The difference in latitude between the two points.
- $\Delta\mathrm{lon}$ - The difference in longitude between the two points.
- $\mathrm{lat}_1, \mathrm{lat}_2$ - The latitudes of the two points being compared.

Within our system, we have established two distinct Vector databases: one for storing photo embeddings and another for face embeddings, each characterized by dimensions of 256 and 512, respectively. As elaborated in Section IV, the vector will be normalized for adding to the database and performing dot product. This process is equivalent to cosine similarity but more efficient.

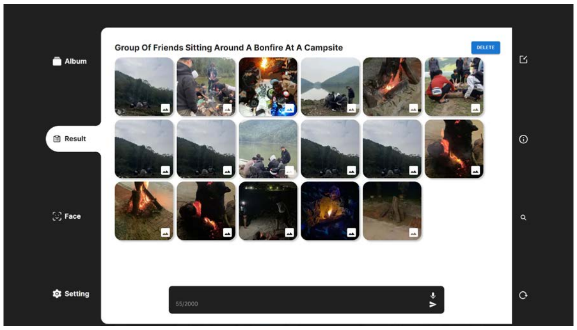
In terms of similarity scoring, we have set a threshold of 0.35 for our dataset. This threshold serves as a benchmark to determine the relevance of search results, ensuring that only highly similar images are retrieved. Additionally, for geospatial data, we have configured a filtering radius (r) of 50 kilometers, enabling users to refine their photo searches based on location with precision.

To optimize the user experience, expedite webpage loading times, reduce the server load, and minimize bandwidth consumption, we have integrated imagekit.io [4] as a third-party service for automatic image and video optimization and delivery. This strategic implementation not only enhances system performance but also significantly improves the efficiency of image and video delivery, thereby enhancing the overall user experience.

C. Use Cases

This section provides several use cases in the Glimpse platform, that demonstrate the completely useful and understandable for use even with general users, who don't have any experience with the platform.

- **Scenario 1:** "Group of friends sitting around a bonfire at a campsite".



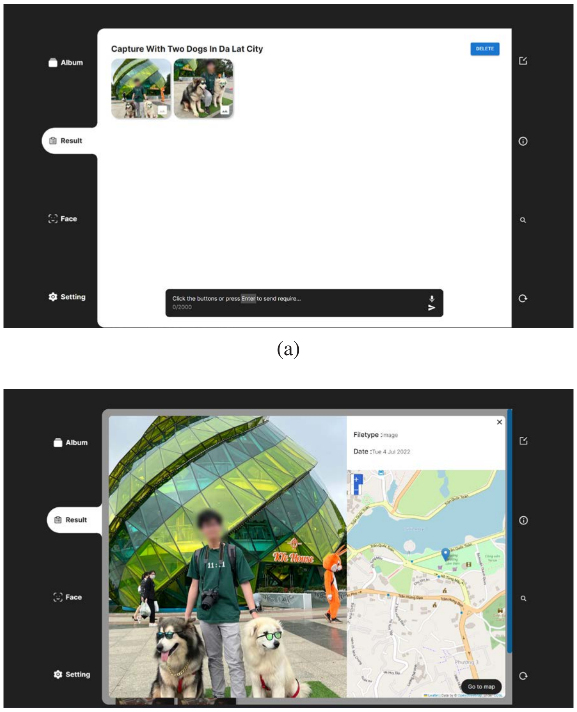
Group Of Friends Sitting Around A Bonfire At A Campsite

55/2000

Fig. 5. Group of friends sitting around a bonfire at a campsite

Within this scenario, the query lacks specific information regarding time, location, or individual names. Consequently, our system conducts a search solely based on the semantic content of the query. The outcome of this search is depicted in Figure 5, where results are generated without the consideration of temporal, spatial, or personal attributes.

- **Scenario 2:** "I'm taking photo with two dogs in Da Lat city".



Capture With Two Dogs In Da Lat City

Click the buttons or press Enter to send require...

0/2000

Filetype :image

Date :Tue 4 Jul 2022

(a)

(b)

Fig. 6. I'm taking photo with two dogs in Da Lat city

In this particular scenario, the query specifies the presence of two dogs and the location as Da Lat city. However, it does not contain any details related to a specific time or individual names. Consequently, our system conducts a search that primarily relies on the semantic content of the query and the location attribute. It omits considerations of temporal information or personal identities. The outcome of this search is visualized in Figure 6, highlighting images depicting the user with two dogs in the setting of Da Lat city, taking into account the query's provided context.

- **Scenario 3:** "Tuong and Tin had dinner together in Ho Chi Minh City last month"

This scenario presents a query that distinctly features two individuals, Tuong and Tin, along with the location, Ho Chi Minh City, and a defined timeframe of "last month". Our system's search approach integrates the semantic essence of the query with the parameters of time, location, and individuals. The illustration of the Named Entity Recognition process for this query is presented in Figure 7. The outcome of this multifaceted search is portrayed in Figure 8, unveiling a collection of images that encapsulate the shared dining experience of Tuong and Tin in Ho Chi Minh City during the specified preceding month.

[4] https://imagekit.io

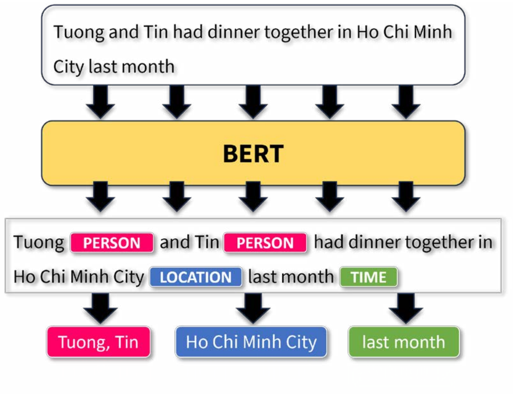
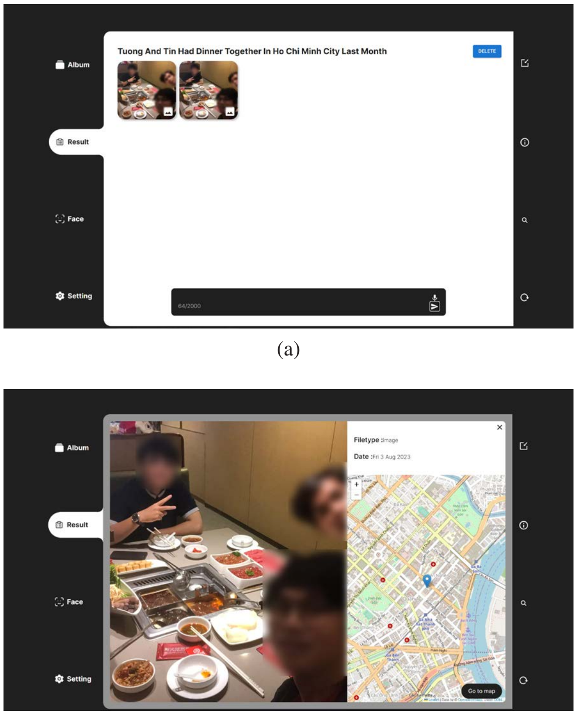

Tuong and Tin had dinner together in Ho Chi Minh City last month
BERT
Tuong PERSON and Tin PERSON had dinner together in Ho Chi Minh City LOCATION last month TIME
Tuong, Tin
Ho Chi Minh City
last month

Fig. 7. The named entity recognition processing in Glimpse


Tuong And Tin Had Dinner Together In Ho Chi Minh City Last Month

(a)

(b)

Fig. 8. Tuong and Tin had dinner together in Ho Chi Minh City last month

VI. Conclusion

In summary, Glimpse, our photo query-based search system, represents an amalgamation of state-of-the-art techniques. Leveraging NER (Named Entity Recognition), Face detection, and BLIP (Bootstrapping Language-Image Pre-training), Glimpse excels in zero-shot image-text matching. Utilizing a vector database for embedding storage, similarity score calculation, and filtering mechanisms ensures the provision of relevant and insightful search results within personal photo albums. In addition, we have developed a user-friendly web application interface, enabling users to swiftly and effortlessly locate images that are not only relevant but also delivered with expeditious ease.

VII. Acknowledgement

This research was supported by AIT Laboratory, FPT University, Danang Campus, Vietnam 2024

References

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic *et al.*, “Query by image and video content: The qbic system,” *computer*, vol. 28, no. 9, pp. 23–32, 1995.

[2] R. Gasser, L. Rossetto, and H. Schuldt, “Multimodal multimedia retrieval with vitrivr,” in *Proceedings of the 2019 on international conference on multimedia retrieval*, 2019, pp. 391–394.

[3] S. Changpinyo, J. Pont-Tuset, V. Ferrari, and R. Soricut, “Telling the what while pointing to the where: Multimodal queries for image retrieval,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 136–12 146.

[4] T. L. Duyen, N. M. Duy, N. T. Binh, H. Lee, and C. Gurrin, “Myscéal-an experimental interactive lifelog retrieval system for lsc’20,” in *Proc. ACM Workshop on Lifelog Search Challenge (LSC@ ICMR 2020). ACM, Dublin, Irelend*, 2020.

[5] N. Alam, Y. Graham, and C. Gurrin, “Memento: A prototype lifelog search engine for lsc’21,” in *Proceedings of the 4th Annual on Lifelog Search Challenge*, 2021, pp. 53–58.

[6] S. Heller, R. Gasser, M. Parian-Scherb, S. Popovic, L. Rossetto, L. Sauter, F. Spiess, and H. Schuldt, “Interactive multimodal lifelog retrieval with vitrivr at lsc 2021,” in *Proceedings of the 4th Annual on Lifelog Search Challenge*, 2021, pp. 35–39.

[7] A. Leibetseder and K. Schoeffmann, “lifexplore at the lifelog search challenge 2021,” in *Proceedings of the 4th Annual on Lifelog Search Challenge*, 2021, pp. 23–28.

[8] D. Sarkar, *Text Analytics with Python: A Practitioner’s Guide to Natural Language Processing*, 01 2019.

[9] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Beijing: O’Reilly, 2009. [Online]. Available: http://www.nltk.org/book

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[11] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[15] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[16] M. Rahman, M. S. Hossain, M. K. Alam, M. E. Kabir, and M. A. Ullah, “Density-based clustering of geo-location data,” *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 1, pp. 17–26, 2017.

[17] R. L. Thorndike, “Who belongs in the family?” *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.