

Exquisitor at the Lifelog Search Challenge 2024: Blending Conversational Search with User Relevance Feedback

Omar Shahbaz Khan
Reykjavik University
Reykjavik, Iceland
omark@ru.is

Ujjwal Sharma
University of Amsterdam
Amsterdam, Netherlands
u.sharma@uva.nl

Hongyi Zhu
University of Amsterdam
Amsterdam, Netherlands
h.zhu@uva.nl

Stevan Rudinac
University of Amsterdam
Amsterdam, Netherlands
s.rudinac@uva.nl

Björn Þór Jónsson
Reykjavik University
Reykjavik, Iceland
bjorn@ru.is

ABSTRACT

The past decade has seen a rapid expansion of personal and interpersonal multimedia collections. These collections offer a wealth of information about individuals, including their interests, health, and significant life events. While automated techniques can assist in structuring and organizing these collections, they often have limitations in helping users effectively navigate and find relevant items within such large datasets. The Lifelog Search Challenge (LSC) provides a valuable benchmark for evaluating interactive retrieval systems designed for personal multimedia collections. Exquisitor utilizes a large-scale user relevance feedback (URF) approach for searching through large collections. To address challenges in highly descriptive retrieval tasks where the relevance feedback model may fail to identify essential elements, we have enhanced Exquisitor with conversational search capabilities powered by a Vision Language Model (VLM) and refined the features underlying the URF model. Furthermore, Exquisitor has been updated with a streamlined user interface that enables seamless switching between conversational search and URF modes.

CCS CONCEPTS

• **Information systems** → **Collaborative search; Users and interactive retrieval.**

KEYWORDS

Lifelogging; Vision Language Models; Interactive learning; Conversational Search; Exquisitor.

ACM Reference Format:

Omar Shahbaz Khan, Ujjwal Sharma, Hongyi Zhu, Stevan Rudinac, and Björn Þór Jónsson. 2024. Exquisitor at the Lifelog Search Challenge 2024: Blending Conversational Search with User Relevance Feedback. In *The 7th Annual ACM Lifelog Search Challenge (LSC '24)*, June 10, 2024, Phuket, Thailand. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3643489.3661132>



This work is licensed under a Creative Commons Attribution-ShareAlike International 4.0 License.
LSC '24, June 10, 2024, Phuket, Thailand
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0550-2/24/06.
<https://doi.org/10.1145/3643489.3661132>

1 INTRODUCTION

Personal multimedia collections have been expanding rapidly in today's society due to the widespread use of smartphones, enabling individuals to capture moments they find noteworthy. While the quantity of data generated and stored by users varies with individuals, its volume has indicated a general upward trend as mobile phones with cameras have become generally ubiquitous. Such extensive collections typically comprise diverse media types including images, videos, audio recordings, and text, along with ancillary metadata from the capturing device such as EXIF tags and geolocation data. Moreover, when smart gadgets are utilized, personal multimedia collections may additionally incorporate biometric data. This data can be further analyzed to provide insights into an individual's mental and physical health status over time or during specific moments. Personal multimedia collections serve multiple purposes, including aiding in the recall of particular events and revealing patterns about oneself. These patterns can encompass a broad spectrum of behaviors and indicators, ranging from mundane habits to critical health-related concerns. Naturally, there is great value in the ability to accurately retrieve visual data corresponding to specific life events or important health markers from a large and growing visual collection.

The Lifelog Search Challenge (LSC) presents a benchmark for testing retrieval abilities associated with visual retrieval systems for searching through and retrieving life events. LSC is a live interactive retrieval competition where participating teams use multimedia retrieval systems to solve tasks related to the retrieval of relevant snapshot(s) from a lifelogger's vast collection [4, 5]. A *lifelogger* is an individual who prioritizes the documentation of extensive personal data throughout their life. In most cases, this involves wearing a miniature camera that captures images at regular intervals, accompanied by geolocation data and metadata from the camera itself. Lifeloggers may further utilize smartwatches to record relevant biometric data. With an image captured every 30 seconds, a lifelogger accumulates thousands of images daily and millions over several years. These extensive collections serve as valuable resources for the lifelogger. Additionally, their volume often necessitates interactive exploration and search tools to facilitate access to desired information. Within LSC, there are three primary types of image retrieval tasks designed around retrieval of lifelogging data:

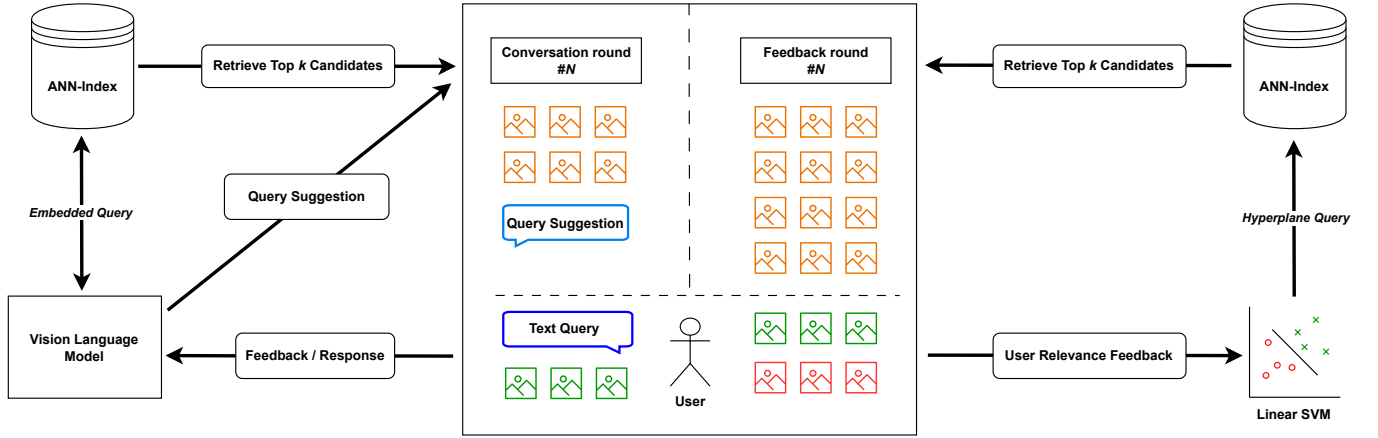


Figure 1: The interactive workflow of Exquisitor. The middle box represents the screen, which splits into conversational search and user relevance feedback. The left-hand side shows the conversational search flow, starting with a user supplying a textual query (along with positive examples if it is not the start of the session). The textual query and the supplied items are processed in the VLM to obtain an embedded query to search with, as well as a query suggestion. The embedded query is used to search an ANN-index to retrieve the top k items (nearest), along with the query suggestion. The right-hand side depicts the user relevance feedback flow, which start by the user labelling positive and negative items. These items are used to train a Linear SVM to create a hyperplane query, which is used to search an ANN-index to retrieve the farthest k items.

- (1) Known Item Search: Finding a specific image related to a known event.
- (2) Ad-hoc Search: Finding as many images as possible that fit a given description.
- (3) Q&A Topics: Providing a textual answer to an information need formulated as a query.

Throughout the development of LSC, promising methods for solving these tasks have emerged. These include temporal-search, object-based search, relevance feedback, and interactive search within virtual reality environments.

Fundamentally, the core functionality of LSC lies in its ability to rapidly search and explore extensive image collections. While various techniques can be used to extract features from multimedia items (facilitating organization and analysis), most people are primarily familiar with two search methods: text-based search and filtering. These approaches enable users to obtain ranked lists of relevant items based on their queries and selected criteria. In cases where a user can describe a specific event and apply suitable filters, the retrieved list is often concise and easily navigable. However, having a clear description is not always possible—events may blend together due to similarity, the details may be hazy for the user, or they may lack the vocabulary for accurate description. This can lead to a cumbersome number of items to sort through. In such situations, a content-based retrieval approach is often advantageous [3, 13]. This can involve similarity search or an iterative feedback process like user relevance feedback. The latter method offers greater flexibility, which is especially useful if the user’s area of interest changes while they are examining the retrieved items.

Exquisitor is a large scale interactive learning approach and system for multimedia collections [7]. At its core it relies on user relevance feedback (URF), where a user labels the retrieved items as positive or negative, which updates the model and retrieves a

new set of items. It uses text-based search to find example items for the feedback process, along with filters that help narrow the search space. Exquisitor has participated in previous iterations of LSC, initially using a pure URF approach along with filters [8], and later with keyword search and temporal querying using multiple feedback models [6, 10]. These participations highlighted the advantages of URF, but also the limitations such a system has for retrieval oriented tasks. URF as the primary tool is not always useful, and using keyword search limits the user to a set of query terms. Furthermore, the interface separated the URF and search into different tabs, leading to users typically using one feature over the other. Having users switch seamlessly between URF and search depending on the details of the tasks is a more desirable interaction flow.

To address these concerns, we have redesigned the interface and added an entirely new conversational search feature to Exquisitor. Figure 1 depicts the new interaction workflow. The screen is split into conversational search on one side, and URF on the other side. The conversational search presents the user with items related to their text query using a Vision Language Model (VLM), and suggests improvements to their query. They can then label positive items along with formulating a new query. The labelled items from the conversational search are also added to the URF model. LSC 2024 is an ideal venue to test how much of a difference these changes make compared to the previous participations.

2 EXQUISITOR

Exquisitor is a multi-paradigm search system featuring two distinct search sub-systems: User relevance feedback driven search and Conversational search.

- (1) User Relevance Feedback driven search (URF): A cornerstone of Exquisitor since its introduction in [8], this pipeline leverages iterative user relevance judgments to refine search

results. This approach is ideal when users have a semantic concept in mind but might struggle to articulate it precisely.

- (2) **Conversational search:** A more recent enhancement [11], conversational search allows users to interact with the collection through intuitive natural language queries. This approach offers a more user-friendly search experience built upon natural language queries, and is generally preferred over URF when the concept to be searched can be distinctively described in language.

Currently, the Exquisitor interface treats these modes of interaction as separate, stand-alone search sub-systems.

2.1 User Relevance Feedback Driven Search

User relevance feedback (URF) is a human-in-the-loop approach that involves presenting a set of suggestions to the user. The initial suggestion set is arbitrarily chosen. Over multiple rounds, the user determines the relevance of the items by labelling them as positives or negatives. The level of feedback requested by the URF system determines the user's task. The user may need to provide feedback on all items in the suggestion set, a fixed number of items, or have the choice left to them. Once their feedback is provided, the labelled items are used to update the system's underlying classification model, such as a Support Vector Machine (SVM) [2]. A new suggestion set, representing the most relevant items of the updated model, is then presented. This feedback-loop interaction continues until the user deems the session over.

In Exquisitor, the classification model in the URF pipeline is a Linear-SVM, that creates a hyperplane from the labeled items. The most relevant items are the ones that are furthest from the decision boundary in the positive direction (maximum inner product). The Linear-SVM only requires a few training samples to be efficient, making it easier to adapt the model on the fly. This makes it ideal for interactive applications where the user information need changes or the user readjusts the feedback on previously labelled items. To enable URF for large scale datasets, Exquisitor uses an approximate high-dimensional index called extended Cluster Pruning (eCP), which is a hierarchical clustering index. A trait of the eCP index is that the clusters have a size limit which keeps them well-balanced, leading to better time guarantees. At runtime, eCP uses the search expansion parameter, b , which is the number of clusters it will check at each level of the tree. Through user tests, it became apparent that b could be detrimental when filters are utilized, since they are applied to the results of the index search. To remedy this, we introduced incremental retrieval along with a set of query optimization policies that dynamically increases b to avoid scenarios of relevant items not being suggested.

2.2 Conversational Search

URF relies on users providing a series of positives and negatives to refine search results over multiple rounds. However, when queries contain clearly defined semantic concepts, this process can be streamlined. For instance, the query "a field with the Icelandic flag in the background" allows for direct visual search based on the "Icelandic flag" – bypassing several URF iterations. Moreover, finding initial positive examples for URF in large collections can be tedious, requiring users to examine random sets, apply filters, and provide

many negative examples early on. To mitigate this, Exquisitor incorporates conversational search, specifically designed for scenarios where the desired image can be described textually. While earlier versions relied on simple ImageNet keyword searches [10, 12], the current system utilizes full text search with a CLIP model [11]. During the search process, users can freely switch between URF-driven search and conversational search. URF excels in exploratory tasks, while conversational search helps locate potential positive examples. Conversational search is effective in retrieval tasks, primarily due to the underlying CLIP model's excellent zero-shot capabilities. URF then aids exploration and provides inspiration to refine the search query [15]. When queries are abstract and lack concrete visual descriptors, URF assumes the leading role in the search process.

2.3 User Interface

The original interface of Exquisitor leaned heavily towards URF [6, 8, 10], with the main part of the screen consisting of a grid of images. A green and red container, containing positives and negatives, were located on the left and right sides of the grid, respectively. Above the grid, a set of filters was available for use. To search for good example items, multiple new tabs could be opened where keyword searches using ImageNet concepts were possible. Although the intention of the interface was to have users predominantly focus on using the URF features, in practice, with highly retrieval-oriented tasks, users typically remained in the search tab. This was especially true for novice users who were unfamiliar with URF, while some forgot to use URF due to it being in a separate tab [9].

3 NEW FEATURES FOR LSC 2024

In this section, we describe the changes introduced to Exquisitor's two search approaches and its interface for LSC 2024.

3.1 User Relevance Feedback

The major change for the URF process of Exquisitor, are the features used in the process. Instead of using semantic concepts from ImageNet, it now uses embeddings from a CLIP model [14]. While the URF process did perform reasonably using the concepts, the top concepts of an item could often be synonyms of each other. This could unintentionally throw the model into completely different directions, losing the trust of users. The move to the CLIP embeddings does present a drastic change, as the CLIP model produces dense vectors compared to the former sparse representation. The sparse nature of the ImageNet concept vectors, made it possible to only use a handful of the top features instead of all, which contributed to the interactive response times of Exquisitor. The same approach cannot be applied to the dense vectors, which have 1368 dimension features, but instead we choose to quantize these features using Principal Component Analysis (PCA) [1]. The resulting feature vectors have 192 dimensions, and while it does reduce accuracy, it is an acceptable trade-off to gain more interactive response times.

3.2 Conversational Search

First introduced in [11], conversational search has proven to be an effective and efficient method to search through collections using natural language queries. This system converts text queries into

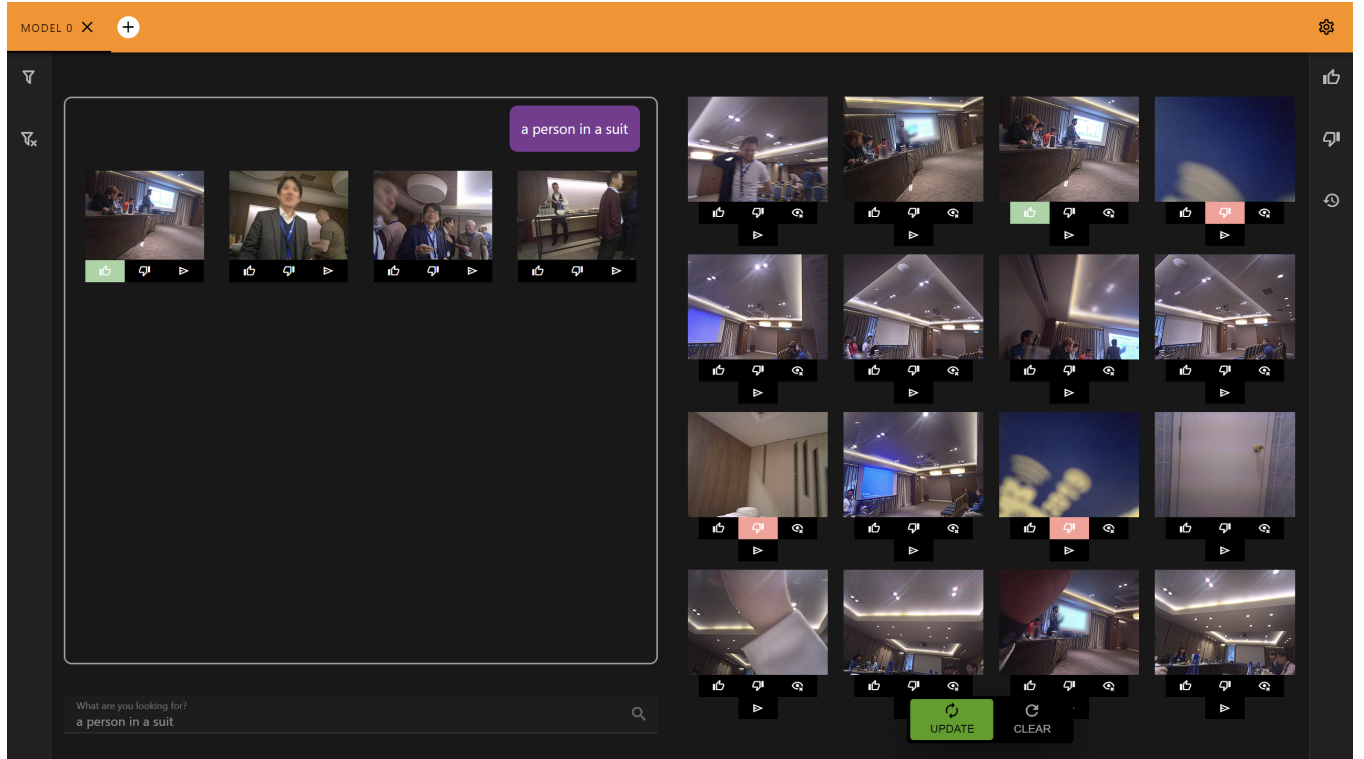


Figure 2: Exquisitor’s combined interactive learning interface. The left-hand side of the interface shows the conversational interface, where a user inputs a query at the bottom and gets a set of suggestions. The right-hand side of the interface shows the user relevance feedback grid where the user can click on the “Like” or “Dislike” buttons to give feedback, (highlighted with green and red respectively when clicked). To request new items from the feedback process, the user clicks the highlighted “Update” button at the bottom.

dense vectors that, coupled with fast approximate nearest neighbor search, allow real-time image search over extremely large image collections. However, the current implementation in Exquisitor is inherently single-shot and stateless – providing results only for a single query at a time. This presents a significant limitation for Exquisitor as it requires users to provide all necessary information within a single query, leading to verbose queries. Moreover, its stateless nature prevents users from refining results over multiple rounds – a capability already supported by Exquisitor’s URF search subsystem. To address this, we have replaced the single-shot search system with a stateful, interactive, and multi-turn approach. This new approach continuously refines search results based on information gathered throughout successive search rounds, aligning the conversational search strategy with the multi-round, interactive nature of Exquisitor’s existing URF search subsystem. The system maintains state and, within each interaction round, refines search parameters using a large vision language model. Our goal with this approach is to enable true conversational search capabilities within Exquisitor and reduce complexity for the end-user.

3.3 Redesigned User Interface

Figure 2 depicts the new fully redesigned user interface for Exquisitor. In this new interface, all features are available on the screen, so

there is no need to navigate to other tabs, unless the user is working with multiple models. The main screen is split into conversational search (left) and URF (right). The filters can be accessed from the buttons on the left sidebar, with the top button being inclusive filters and the bottom being exclusive filters (if applicable). On the right sidebar the user can access the current positive, negative, and previously seen items in the current URF session.

4 CONCLUSION

In this paper, we describe the multi-paradigm search system Exquisitor. Exquisitor allows users to use conversational search for interaction in a familiar chat-based manner, and user relevance feedback to explore and search the collection in a more conventional content-based manner. The user interface of Exquisitor has also been redesigned from its previous participations in LSC, where it leaned primarily towards URF, with the aim of both search approaches being easy to use for all types of users.

ACKNOWLEDGMENTS

This work was partially supported by Icelandic Research Fund (IRF) grant 239772-051.

REFERENCES

- [1] Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* 2, 4 (2010), 433–459.
- [2] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20, 3 (1995), 273–297.
- [3] Shiv Ram Dubey. 2021. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 5 (2021), 2687–2704.
- [4] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Duc-Tien Dang-Nguyen, Michael Riegler, and Luca Piras (Eds.). 2018. *Proceedings of the ACM Workshop on Lifelog Search Challenge, LSC 2018*. ACM, Yokohama, Japan.
- [5] Cathal Gurrin, Liting Zhou, Graham Healy, Werner Bailer, Duc-Tien Dang-Nguyen, Steve Hodges, Björn Þór Jónsson, Jakub Lokoč, Luca Rossetto, Minh-Triet Tran, and Klaus Schöffmann. 2024. Introduction to the Seventh Annual Lifelog Search Challenge, LSC'24. *International Conference on Multimedia Retrieval (ICMR'24)*. <https://doi.org/10.1145/3652583.3658891>
- [6] Omar Shahbaz Khan, Aaron Duane, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, and Marcel Worring. 2021. Exquisitor at the Lifelog Search Challenge 2021: Relationships Between Semantic Classifiers. In *Proceedings of the 4th Annual on Lifelog Search Challenge (LSC '21)*. Association for Computing Machinery, New York, NY, USA, 3–6.
- [7] Omar Shahbaz Khan, Björn Þór Jónsson, Stevan Rudinac, Jan Zahálka, Hanna Ragnarsdóttir, Þórhildur Þorleiksdóttir, Gylfi Þór Guðmundsson, Laurent Amsaleg, and Marcel Worring. 2020. Interactive Learning for Multimedia at Large. In *Proceedings of the European Conference on Information Retrieval (ECIR)*. Springer, Lisboa, Portugal, 16.
- [8] Omar Shahbaz Khan, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, and Marcel Worring. 2019. Exquisitor at the Lifelog Search Challenge 2019. In *Proceedings of the ACM Workshop on Lifelog Search Challenge, LSC 2019*. ACM, Ottawa, ON, Canada, 7–11.
- [9] Omar Shahbaz Khan and Björn Þór Jónsson. 2023. User Relevance Feedback and Novices: Anecdotes from Exquisitor's Participation in Interactive Retrieval Competitions. In *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing (CBMI '23)*. Association for Computing Machinery, New York, NY, USA, 173–177.
- [10] Omar Shahbaz Khan, Mathias Dybkjær Larsen, Liam Alex Sonto Poulsen, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, Dennis Koelma, and Marcel Worring. 2020. Exquisitor at the Lifelog Search Challenge 2020. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*. ACM.
- [11] Omar Shahbaz Khan, Hongyi Zhu, Ujjwal Sharma, Evangelos Kanoulas, Stevan Rudinac, and Björn Þór Jónsson. 2024. Exquisitor at Video Browser Showdown 2024: Relevance Feedback Meets Conversational Search. In *MultiMedia Modeling: 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 – February 2, 2024, Proceedings, Part IV*. Springer-Verlag, Berlin, Heidelberg, 347–355.
- [12] Pascal Mettes, Dennis C Koelma, and Cees GM Snoek. 2016. The imagenet shuffle: Reorganized pre-training for video event detection. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. 175–182.
- [13] Arnold WM Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence* 22, 12 (2000), 1349–1380.
- [14] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11975–11986.
- [15] Hongyi Zhu, Jia-Hong Huang, Stevan Rudinac, and Evangelos Kanoulas. 2024. Enhancing Interactive Image Retrieval With Query Rewriting Using Large Language Models and Vision Language Models. *arXiv e-prints*, Article arXiv:2404.18746 (April 2024), arXiv:2404.18746 pages. arXiv:cs.MM/2404.18746