# An Exploration into the Benefits of the CLIP model for Lifelog Retrieval

Ly-Duyen Tran*
Naushad Alam*
ly.tran2@mail.dcu.ie
naushad.alam2@mail.dcu.ie
Dublin City University
Dublin, Ireland

Linh Khanh Vo
Nghiem Tuong Diep
AISIA Research Lab
Ho Chi Minh, Vietnam
University of Science
Ho Chi Minh, Vietnam
Vietnam National University
Ho Chi Minh, Vietnam

Binh Nguyen
AISIA Research Lab
Ho Chi Minh, Vietnam
University of Science
Ho Chi Minh, Vietnam
Vietnam National University
Ho Chi Minh, Vietnam

Yvette Graham
Trinity College
Dublin, Ireland

Liting Zhou
Dublin City University
Dublin, Ireland

Cathal Gurrin
Dublin City University
Dublin, Ireland

## ABSTRACT

In this paper, we attempt to fine-tune the CLIP (Contrastive Language-Image Pre-Training) model on the Lifelog Question Answering dataset (LLQA) to investigate retrieval performance of the fine-tuned model over the zero-shot baseline model. We train the model adopting a weight space ensembling approach using a modified loss function to take into account the differences in our dataset (LLQA) when compared with the dataset the CLIP model was originally pretrained on. We further evaluate our fine-tuned model using visual as well as multimodal queries on multiple retrieval tasks, demonstrating improved performance over the zero-shot baseline model.

## CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; • **Information systems** → **Information retrieval**.

## KEYWORDS

lifelogging, image retrieval, pretrained models

**ACM Reference Format:**
Ly-Duyen Tran, Naushad Alam, Linh Khanh Vo, Nghiem Tuong Diep, Binh Nguyen, Yvette Graham, Liting Zhou, and Cathal Gurrin. 2022. An Exploration into the Benefits of the CLIP model for Lifelog Retrieval. In *International Conference on Content-based Multimedia Indexing (CBMI 2022), September 14–16, 2022, Graz, Austria.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3549555.3549593

## 1 INTRODUCTION

*"Where are my car keys?"*. Human memory can be fallible and unreliable, though it is undoubtedly provides a vital cognitive ability 19. Humans tend to constantly forget trivial things, such as failing to remember the location of items, details about a recent event, or simply struggling to remember the name of the person we just met. In this work, we are interested in supporting human memory by building a digital twin of an individual that can answer all such daily information needs.

Lifelogging, as defined by 17, is the process of passively capturing a personal digital collection of daily life experiences using a variety of devices, such as wearable cameras, tracking devices such as FitBit, and other wearable sensor devices. As a concept, lifelogging was introduced in Vannevar Bush's 1945 article 'As We May Think' 4 where he discusses about a "future mechanised device" which acts as an "enlarged intimate supplement of an individual's memory" storing all his books, records, communications and can be consulted with "exceeding speed and flexibility".

The last two decades have witnessed growing attention to lifelogging after MyLifeBits 13 was proposed by Gemmell and Bell in the early 2000s. Furthermore, with advances in sensor technology, the availability of cheap storage facilities and cost-efficient wearable devices, recording one's life passively has become feasible, and hence lifelogging has witnessed a surge in interest from the research community. Due to the sheer variety of data collected, lifelog data have been used to address various use cases in research domains such as signal processing 9, 39, natural language processing 26, 53, computer vision 50, 53, and human-computer interactions 49.

Information retrieval from lifelogs to realise the goal of memory augmentation is, however, a very challenging problem as human memory is pervasive and immediate while retrieval from lifelogs using implicit queries is an iterative and cumbersome process. The multimodal characteristics of lifelog data, which includes data from multiple sources such as egocentric images, textual data specifying details like location, time, date and biometrics as well as it being a noisy and repetitive archive due to passive data collection over longer periods of time, further adds to the challenges of developing an effective retrieval system.

[*] Both authors contributed equally to this research.

*CBMI 2022, September 14–16, 2022, Graz, Austria*

ACM ISBN 978-1-4503-9720-9/22/09.
https://doi.org/10.1145/3549555.3549593

Recent models like Contrastive Language-Image Pre-Training (CLIP) 36, A Large-scale ImaGe and Noisy-text embedding (ALIGN) 24, etc. which leverage the supervision inherent in natural language texts to learn generalised vision-language representations, can be used to solve multiple downstream tasks such as information retrieval, object recognition, scene recognition. These have seen tremendous success on multiple benchmarks recently. The zero-shot CLIP model 36 beats several supervised task-specific models on multiple datasets, showing robust transfer capability when applied to out-of-domain datasets. However, as discussed in Section 2.2 the model fares poorly when applied to certain specialised datasets, which motivated us to experiment with model fine-tuning on lifelogs and compare the performance the zero-shot model.

This work aims to investigate whether fine-tuning the CLIP model on domain-specific data improves the model's performance when compared with the zero-shot baseline model to solve the task of lifelog information retrieval. Our contributions in this paper are as follows.

- We fine-tune the CLIP model on an in-the-wild egocentric multimodal dataset (Lifelogs), which to the best of our knowledge is the first work done in this domain.
- We devise a modified loss function to accommodate the structure of the dataset we use to fine-tune the CLIP model.
- We evaluate our fine-tuned model on multiple retrieval tasks using both visual and multimodal queries demonstrating superior performance over the zero-shot baseline model.

The results of this study will provide insights into the design decisions of lifelog retrieval systems in the future. This paper is structured as follows: Section 2 discusses the efforts carried out so far to fine-tune the CLIP models on various niche datasets, as well as briefly covering the major milestones achieved so far at a high level in the area of transfer learning. Subsequently, Section 3 discusses the question answering LLQA dataset which has been used to fine-tune the CLIP model followed by a detailed discussion on our adopted methodology. Finally, in Section 4, we evaluate the performance of our fine-tuned model over the zero-shot baseline using both visual queries as well as multimodal queries as input.

## 2 RELATED WORK

### 2.1 Transfer Learning

Curating a high-quality, large-scale annotated dataset is a challenge in many specialised research domains, making it hard to train large deep learning models from scratch. Transfer learning aims to solve this issue of insufficient training data by 'transferring' knowledge from a data-abundant source domain to a data-scarce target domain 50. For a long time now, researchers have used pretrained features from ImageNet 7 to solve several downstream computer vision tasks such as image classification, object detection, action recognition, image segmentation, etc. In recent years, transfer learning in the form of pretrained language models 8 31 38 based on the tranformer architecture 45 has become quite ubiquitous in the field of natural language processing as well, achieving state-of-the-art performance in areas like machine translation, natural language inference, etc.

Recently, models like CLIP 36 and ALIGN 24 generating zero-shot transferable representations have made a leap forward towards generalised models which can work without any data-specific fine-tuning. However, as discussed in Section 2.2, zero-shot transfer to few specific domains is still very challenging. Consequently, several recent works have tried to leverage the pretrained CLIP model to further improve its performance by fine-tuning it on various specialised datasets and have demonstrated competitive performance over the zero-shot model.

### 2.2 Fine-tuning CLIP

As discussed in 36, the zero-shot CLIP model has shown significant gains over the performance of fully supervised ResNet-50 20 baselines trained on several datasets. However, the zero-shot model fails to surpass the performance of supervised models on a few specialised datasets such as EuroSAT 21 and RESISC45 5 (satellite image classification), PatchCamelyon 46 (lymph node tumour detection), CLEVRCounts 25 (counting objects in synthetic scenes), GTSRB 41 (German traffic sign recognition), KITTI Distance 12 (recognising distance to the nearest car), which shows the room for improvement for such complex and niche datasets.

Several recent works have tried to fine-tune the network to improve its performance over specialised datasets in order to address abstract problems. Clip-Art 6 aims to solve the problem of retrieving and classifying fine-grained attributes of artwork images by fine-tuning the network on the iMet dataset. PointCLIP 52 fine-tuned the model with the objective of transfer learning across different modalities. It learns to efficiently transfer representations learnt from 2D images to do cross-modality zero-shot recognition on a 3D point cloud. Arutiunian et al. 3 fine-tuned the model on satellite images and captions from the RSICD dataset 35 to support satellite image retrieval using queries in natural language. In addition, ActionCLIP 47 applied the model to perform video action recognition.

Another line of work focusses on strategies and techniques to robustly fine-tune the CLIP model. CLIP Adapter 11 adopts a light-weight bottleneck architecture to prevent the potential overfitting problem of few-shot learning by reducing the number of parameters and only fine-tuning a small number of additional weights instead of optimising all CLIP parameters. Tip-Adapter 51 (Training free CLIP-Adapter) further improves over CLIP-Adapter 11 by doing away with stochastic gradient descent to train the adapter and instead constructing a query-key cache model from few-shot supervisions to obtain the weights of the adapter. WiSE-FT 48 proposed to do weight space ensembling between zero-shot and fine-tune models to preserve the model's accuracy under data distribution shift.

In this work, we attempt to fine-tune the CLIP model on the lifelog dataset, which is an in-the-wild egocentric multimodal dataset adopting the weight-space ensembling approch from 48 demonstrating encouraging results.

### 2.3 Lifelog Retrieval

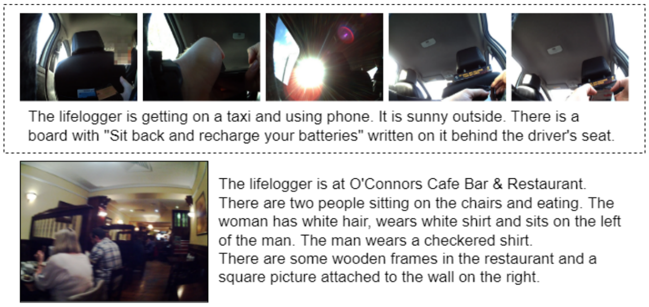
Effective information retrieval from lifelogs has been a longstanding challenge given the multimodal nature and size of the dataset as well as the very specific nature of information users would want to retrieve from it. In recent years, several benchmarking challenges have been organised to support collaborative benchmarking for lifelog retrieval systems, one example of which is the Lifelog Search Challenge (LSC) 15.

The LSC workshop has attracted many interactive lifelog retrieval systems that are based on existing video retrieval systems that regularly participate in the similar Video Browser Showdown (VBS) Challenge 22, such as 23, 29, 32. There have been a number of novel approaches to address the lifelog challenge, such as Exquisitor 27 which uses a dynamic semantic classifier, THUIR 30 which incorporates relevance feedback to guide the search process, and some virtual reality systems, such as 10, 40 which provide a fully immersive experience to the search process.

Likewise, several past systems leveraged visual concepts derived from object detection models to build their retrieval engines 2, 37, 44. Additionally, we note the recent popularity of approaches that take advantage of multimodal embeddings (such as the CLIP zero-shot model) 53 for lifelog retrieval, which potentially have surpassed prior state-of-the-art techniques in the field 1, 23, 43.

# 3 EXPERIMENT SETUPS

## 3.1 Lifelog captions



The lifelogger is getting on a taxi and using phone. It is sunny outside. There is a board with "Sit back and recharge your batteries" written on it behind the driver's seat.

The lifelogger is at O'Connors Cafe Bar & Restaurant. There are two people sitting on the chairs and eating. The woman has white hair, wears white shirt and sits on the left of the man. The man wears a checkered shirt. There are some wooden frames in the restaurant and a square picture attached to the wall on the right.

**Figure 1: Examples of annotated descriptions in the updated LLQA dataset. The first example feature general narrative descriptions for a longer episode of activity. The second one include details in a single image, usually when the lifelogger was moving and the surroundings change considerably. Each sentence in these descriptions count as one caption in the LLQA dataset used to fine-tune CLIP models.**

**Table 1: Statistics of the number of images that one caption describes.**

| | #captions | mean | min | 25th | 50th | 75th | max |
|---|---|---|---|---|---|---|---|
| Original | 11398 | 15.89 | 1 | 5 | 8 | 15 | 297 |
| Updated | 1919 | 3.39 | 1 | 1 | 1 | 1 | 112 |

To generate a dataset to fine-tune the CLIP models, we utilise the Lifelog Question Answering Dataset (LLQA) 42, which includes questions and answers automatically generated from human-annotated captions. Although LLQA was not created for the purpose of this paper, the number of lifelog captions collected is the largest to the best of our knowledge. A total of 11,398 captions are available in the dataset, describing daily activities from within 85 days of lifelog. However, since this work was an initial attempt at lifelog captioning and question answering, a large portion of the descriptions are vague and not specific enough for some use cases (as previously seen in some queries in the Lifelog Search Challenges (LSC) 15, 16, 18). For this reason, since the publication 42, we have added more captions to the original dataset, including more detailed descriptions that are more suitable. An addition of 1,919 captions are added, which describe details of each image instead of general activities. The comparison between these two parts of the dataset is presented in Table 1. Furthermore, Figure 1 shows two examples from the dataset.

One challenge in adapting this dataset is that the description can describe a period of activity, including multiple images, some of which do not match the caption individually. Thus, we filter the dataset to choose only instances where the caption covers at most 15 images to reduce the possibility of ill-matched pairs of caption and image. The chosen instances are then divided into a training set and a validating set, as detailed in Table 2 to fine-tune the CLIP models, making sure there are no overlapping images between the two sets.

**Table 2: LLQA dataset splits to fine-tune CLIP models.**

| Split | #image-caption pair | #unique images | #unique captions |
|---|---|---|---|
| Train | 11982 | 6328 | 2916 |
| Validate | 1234 | 421 | 328 |

## 3.2 Fine-tuning CLIP models

CLIP models 36 were originally designed to match a single image with a single caption. However, since the descriptions we use often span across multiple images, we modify the loss function accordingly to take into account this characteristic. For every mini-batch, with $T$ as text embedding matrix, $I$ as image embedding matrix, we calculate text similarity $S_T$ and image similarity $S_I$ using the cosine similarity function in Equation 5. The pairwise similarities between text and image, which are $Logits$, are aimed to match with the mean self similarity (of text and image) $Target$ using cross-entropy loss,

$$\mathrm{S_T} = \frac{\mathbf{T} \cdot \mathbf{T}^{\tau}}{||\mathbf{T}||||\mathbf{T}^{\tau}||};\ \mathrm{S_I} = \frac{\mathbf{I} \cdot \mathbf{I}^{\tau}}{||\mathbf{I}||||\mathbf{I}^{\tau}||} \tag{1}$$

$$Logits = \frac{\mathbf{T} \cdot \mathbf{I}^{\tau}}{||\mathbf{T}||||\mathbf{I}^{\tau}||};\ Target = \sigma\left(c \cdot \frac{\mathrm{S_T} + \mathrm{S_I}}{2}\right) \tag{2}$$

$$Loss = crossEntropy(Logits, Target) \tag{3}$$

where $\sigma$ is the softmax function and $c$ is the logit scale.

Due to our limitation of GPU power, we could not fine-tune the best performing model, 'ViT-L/14', amongst the public releases. Instead, we choose to use the pretrained 'ViT-B/32' and 'ViT-B/16' for our experiments. In order to prevent overfitting, we selected the largest minibatch size possible on our machine, which is 48 and 24 for the two models, respectively. We use Adam Optimiser28 with a weight decay regularization34 of 0.01, except for gains or biases, and decay the learning rate using a cosine scheduler 33.

Table 3: Some examples of the tasks in LSC’21 and the corresponding inputs for the CLIP models.

| Task ID | Original Hints | Transformed Hints |
|---|---|---|
| 4 | (1) I needed to buy a blood pressure monitor. (2) So I was looking in a pharmacy (3) that sold Omron and Braun devices. (4) Afterwards, I waited for a long time in my dentist office, (5) before getting a coffee/bagel and driving to my own office. (6) It was in 2016. | (1) Looking to buy a blood pressure monitor (2) in a pharmacy (3) that sold Omron and Braun devices. |
| 18 | (1) I was looking at small computer chips on rolls. (2) It was in a small university electronics laboratory in China. (3) There were at least 100 rolls of small computer chips. (4) It was part of a tour of computing and engineering facilities (5) and I was with a small delegation of people. (6) It was in May 2018. | (1) Looking at small computer chips on rolls (2) in a small university electronics laboratory (3) which had at least 100 rolls. |

Large pretrained CLIP models can perform zero-shot inference with consistent accuracy across a variety of data, which is a valuable characteristic that we want to maintain. For this reason, Wortsman et al. 48 suggested the idea of interpolating the weights between the fine-tuned model and the original to improve robustness. In other words, the final weights of the model are as follows.

$$\theta_{\text{final}} = (1 - \alpha) \cdot \theta_{\text{original}} + \alpha \cdot \theta_{\text{fine-tuned}} \tag{4}$$

The authors suggest choosing $\alpha = 0.5$ as it produced near optimal performance in various experiments. More details on how $\alpha$ affects the performance of the models are provided in Section 4.

## 4 EVALUATION AND RESULTS

Despite the fact that a large proportion of current lifelog data are images, lifelog data are intrinsically multimodal. As CLIP models are incapable of explicit information, such as time or date, in this section, we will adopt CLIP models in two ways:

- **Image-only**: we simplify the queries and include only content-based description;
- **Multimodal**: we incorporate CLIP model with query parsing, automatically extract non-visual information from the query and apply corresponding other search operations.

Two metrics are used to evaluate the models:

- Hit rate at K (*H@K*): *H@K* = 1 means that one of the target images appears in the top K of the result set. Otherwise, *H@K* = 0;
- Average Precision (*AP@K*): the mean of the precision scores after each relevant document is retrieved, where $K$ is the total of relevant documents.

### 4.1 Image-only

At the time of writing, the most recent iteration of the Lifelog Search Challenge, LSC’2115, presented a total of 23 queries with various difficulty levels. Each query was gradually revealed over numberous 30-second time intervals, showing more hints of visual descriptions, time, location, etc. at each interval. For this experiment, we simplify the queries and consider only three time steps, in a way similar to the approach in 1. Some examples of the original queries and the simplified hints are shown in Table 3.

For each query, we encode the query using the textual encoder of the CLIP model and calculate the similarity score of each image with the text embedding. The images are then ranked on the basis of their similarity score. With $\mathbf{q}$ as the encoded search query, and $\mathbf{c}$ as the encoded image, the similarity is defined as:

$$\cos(\mathbf{q}, \mathbf{c}) = \frac{\mathbf{qc}}{\|\mathbf{q}\|\|\mathbf{c}\|} = \frac{\sum_{i=1}^{n} \mathbf{q}_i \mathbf{c}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{q}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{c}_i)^2}} \tag{5}$$

The task of LSC is to find *one* instance of the lifelog moment that matches the search query. Thus, we use the Hit rate at K to measure the performance on these queries. The performance of the public release version of ViT-B/16 can be seen in Table 5, which shows the average *H@K* of each time step.

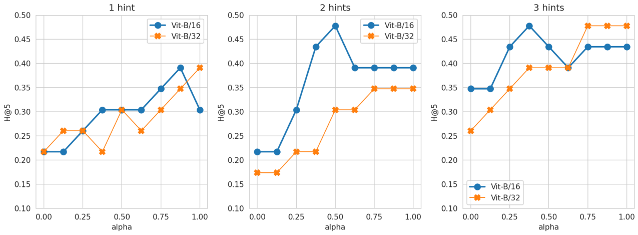


Figure 2: Affect of $\alpha$ on H@5 when interpolation the fine-tuned models.

As mentioned in the previous section, we assemble the fine-tuned CLIP model with the original pretrained weights. To choose the best value for the interpolation parameter $\alpha$, we recorded H@5 scores across all interpolated models to assess the influence of $\alpha$. In general, the fine-tuned models increased the retrieval result after fine-tuning, as can be seen when $\alpha = 1.0$. For the Vit-B/32 model, the increase in H@5 is mostly positively correlated with alpha. However, the pattern for the interpolated weights for ViT-B/16 models is less definite. However, the H@5 scores tend to be higher around the middle point when using two and three hints. For this reason, with the original suggestion from 48, from this point on, we choose to evaluate the fine-tuned ViT-B/16 with $\alpha = 0.5$ on different tasks and address it as **LifelogCLIP** for the sake of simplicity.

The performance of LifelogCLIP is detailed in Table 6. The table shows an increase in almost all hit rates, compared to the original result in Table 5. Surprisingly, the H@1 score for $h = 2$ is lower

**Table 4: Inputs of the same tasks in Table 3 for LifelogCLIP in E-MyScéal. * indicates omitting information due to the limitation of the system.**

| Task ID | Before | Main | After |
|---|---|---|---|
| 4 | | Buying a blood pressure monitor in a pharmacy that sold "Omron" and "Braun" device in 2016. | I waited for a long time in my dentist office.* |
| 18 | A tour of computing and engineering facilities and I was with a small delegation of people | I was looking at small computer chips on rolls. There were at least a hundred rolls of small computer chips. It was in a small university electronics laboratory in China in May 2018 | |

**Table 5: Original CLIP ViT-B/16 performance on 23 queries of LSC’21**

| | H@1 | H@3 | H@5 | H@10 | H@20 | H@50 | H@100 |
|---|---|---|---|---|---|---|---|
| h=1 | 0.17 | 0.22 | 0.22 | 0.26 | 0.35 | 0.48 | 0.52 |
| h=2 | 0.17 | 0.22 | 0.22 | 0.35 | 0.43 | 0.52 | 0.52 |
| h=3 | 0.26 | 0.30 | 0.35 | 0.35 | 0.48 | 0.57 | 0.61 |

**Table 6: LifelogCLIP (fine-tuned CLIP ViT-B/16 with $\alpha = 0.5$) performance on 23 queries of LSC’21**

| | H@1 | H@3 | H@5 | H@10 | H@20 | H@50 | H@100 |
|---|---|---|---|---|---|---|---|
| h=1 | 0.26 | 0.30 | 0.30 | 0.35 | 0.35 | 0.52 | 0.65 |
| h=2 | 0.21 | 0.35 | 0.43 | 0.48 | 0.48 | 0.57 | 0.65 |
| h=3 | 0.30 | 0.43 | 0.48 | 0.48 | 0.52 | 0.61 | 0.65 |

than that of $h = 1$, considering the intuitive assumption that more hints should increase the score, as seen in other cases. This can be explained by the fact that CLIP classifiers can be sensitive to wording or phrasing 36. Hence, adding more information, which changes the phrasing, does not always improve the performance. Other than that, the most significant improvements tend to be in the first row where $h = 1$ and in lower values $K$. This proves that the fine-tuned model more effective in ranking the results.

## 4.2 Multimodal

Taking into account temporal and spatial clues, we incorporate the LifelogCLIP model with the query parsing unit from E-MyScéal 43, a state-of-the-art interactive lifelog retrieval system. To facilitate free-text querying (as opposed to using multimodal faceted filters), the query parsing unit detects location names, as well as date and time format using part-of-speech tagging, semantic role labelling, and regex matching.

*4.2.1 LSC’21 queries.* LSC’21 queries are complex and usually include multiple temporal-related events. To address these queries, the user interface of E-MyScéal accepts *up to* three temporal hints as ‘before’, ‘main’, and ‘after’ queries to address the temporal context in the original search query. Thus, we manually split the LSC’21 queries into temporal queries if needed. Examples of transformed hints are shown in Table 4. Note that in some cases, there are more than one ‘before’ events or more than one ‘after’ events. Due to the limitation of E-MyScéal, we only use the first event in order of appearance. For example, the ’before’ clue ‘*getting a coffee/bagel and driving to my own office*’ in Task ID 4 (Table 3) is left out.

**Table 7: Mean $H@K$ for LSC’21 queries, using all hints.**

| H@1 | H@3 | H@5 | H@10 | H@20 | H@50 | H@100 |
|---|---|---|---|---|---|---|
| 0.52 | 0.65 | 0.74 | 0.74 | 0.74 | 0.74 | 0.74 |

Using all hints, E-MyScéal’s LifelogCLIP can find the answer to more than half of the queries in the first result as seen in Table 7. Interestingly, there is no difference in the hit rates when $k \geq 10$. Since the interface of E-MyScéal can accommodate 12 events at once, this minimises the user’s effort to scroll further down the result page.

*4.2.2 Comparing with baselines on NTCIR-13 lifelog queries.* Since LSC’21 queries are aimed at retrieving a specific moment in lifelog data, hit rate is a suitable metric for evaluation. We also want to assess LifelogCLIP for a different type of lifelog retrieval task with another conventional information retrieval metric: average precision (AP@K). About half of the queries in the NTCIR-13 lifelog 14 challenge focus on retrieving many instances of an activity. We choose the first ten queries and compare LifelogCLIP with the reported performance of two state-of-the-art embedding models from 53. Similarly, we also used the cut-off point at 10 to calculate AP.

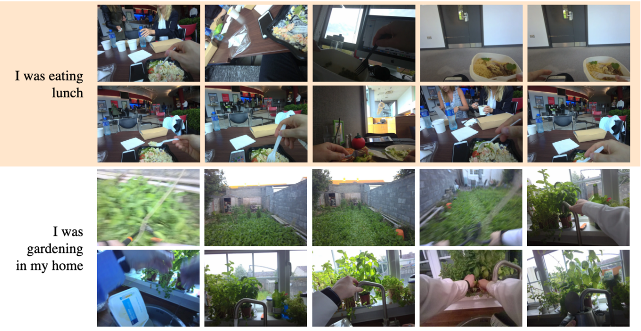


**Figure 3: Top 10 retrieved result from the first two tasks in NTCIR-13.**

Table 8: The AP@10 evaluated on the first 10 tasks of NTCIR-13, compared to the baseline approaches in 53

| Task | Description | Caption | Joint embedding | LifelogCLIP |
|---|---|---|---|---|
| 1 | Find the moments when I was eating lunch | 0.65 | 0.88 | **0.88** |
| 2 | Find moments when I was gardening in my home | 0.12 | 0.23 | **0.40** |
| 3 | Find the moment when I was visiting a castle at night | 0.51 | 0.67 | **0.78** |
| 4 | Find the moments when I was drinking coffee in a cafe | 0.60 | 0.70 | **0.88** |
| 5 | Find the moments when I was outside at sunset | 0.56 | **0.64** | 0.51 |
| 6 | Find the moments when I visited a graveyard | 0.54 | 0.43 | **1.00** |
| 7 | Find the moments when I was lecturing to a group of peoplein a classroom environment | 0.35 | 0.55 | **0.58** |
| 8 | Find all the moments when I was grocery shopping | 0.62 | 0.68 | **1.00** |
| 9 | Find the moments when I worked at home late at night | 0.67 | **0.71** | 0.66 |
| 10 | Find the moments when I was working on the computer at my office desk | 0.57 | 0.85 | **1.00** |

Table 8 details the task descriptions and the performance of the baseline models and LifelogCLIP. For each description, we remove the first part of *"Find the moment when"* or similar phraseing and use only the action (*"I was eating lunch"*) as the search query. As we can see from the table, LifelogCLIP achieved a higher score on most tasks, an equal score on one task, and a lower score on two tasks. Figure 3 illustrates the retrieval results using LifelogCLIP with query parsing. Since we are using LifelogCLIP on an image level, several results in the figure belong to the same event cluster.

The automatic retrieval results of multimodal LifelogCLIP on LSC'21 queries and NTCIR-13 lifelog queries demonstrate that the incorporation of CLIP models can increase the performance of lifelog moment retrieval on both metrics that we proposed at the beginning of the section.

## 5 CONCLUSION

This paper has described our efforts to fine-tune the CLIP models by collecting annotated lifelog descriptions, modifying a loss function for fine-tuning, and evaluating the fine-tuned model on different lifelog retrieval tasks. Its performance is also compared with the baseline multimodal embedding models for lifelog. In summary, we have obtained encouraging results, demonstrating that integrating the fine-tuned CLIP model with query parsing can comparatively enhance the retrieval performance. However, some limitations should be considered. First, the LLQA 42 dataset used for fine-tuning, despite being the best free-form collection of lifelog, is not in the format where CLIP models are usually trained (i.e. having exact matching image-caption pairs). Second, the dataset is small in size for a deep learning task and might not have introduced enough difference for the model to better adapt to lifelog data. Lastly, CLIP models are incapable of taking into consideration the temporal aspect of lifelog. For example, there might be details in a query that while describing the same lifelog moment, they do not appear in a single image. Such queries could lower the retrieval performance, the more detailed they are. More studies are needed to explore these points. In particular, research on solving the last point is already in progress. Additionally, CLIP models, especially the more powerful pretrained versions, are being integrated in more systems in the next Lifelog Search Challenge 18 in various approaches. This provides us a great oppurtunity to ascertain the performance of CLIP models in the future.

## ACKNOWLEDGMENTS

This publication has emanated from research supported in party by research grants from Science Foundation Ireland under grant numbers SFI/12/RC/2289, SFI/12/RC/2289-P2, SFI/13/RC/2106, 18/CRT/6223 and 18/CRT/6224.

## REFERENCES

[1] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2021. Memento: A Prototype Lifelog Search Engine for LSC'21. In *Proceedings of the 4th Annual on Lifelog Search Challenge*. 53–58.

[2] Wei-Hong Ang, An-Zi Yen, Tai-Te Chu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. LifeConcept: An Interactive Approach for Multimodal Lifelog Retrieval through Concept Recommendation. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 47–51. https://doi.org/10.1145/3463948.3469070

[3] Artashes Arutiunian, Dev Vidhani, Goutham Venkatesh, Mayank Bhaskar, Ritobrata Ghosh, and Sujit Pal. 2021. Fine tuning CLIP with Remote Sensing (Satellite) images and captions. https://huggingface.co/blog/fine-tune-clip-rsicd.

[4] Vannevar Bush. 1945. As We May Think. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/ Section: Technology.

[5] Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* 105, 10 (Oct. 2017), 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998 Conference Name: Proceedings of the IEEE.

[6] Marcos V. Conde and Kerem Turgutlu. 2021. CLIP-Art: Contrastive Pre-training for Fine-Grained Art Classification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Nashville, TN, USA, 3951–3955. https://doi.org/10.1109/CVPRW53098.2021.00444

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Chelsea Dobbins and Stephen Fairclough. 2019. Signal Processing of Multimodal Mobile Lifelogging Data Towards Detecting Stress in Real-World Driving. *IEEE Transactions on Mobile Computing* 18, 3 (2019), 632–644.

[10] Aaron Duane and Bjorn Þór Jónsson. 2021. ViRMA: Virtual Reality Multimedia Analytics at LSC 2021. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 29–34. https://doi.org/10.1145/3463948.3469067

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *arXiv:2110.04544 [cs]* (Oct. 2021). http://arxiv.org/abs/2110.04544 arXiv: 2110.04544.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)* (2013).

[13] Jim Gemmell, Chester Bell, and Roger Lueder. 2006. MyLifeBits: A personal database for everything. *Commun. ACM* 49 (01 2006), 89–95.

[14] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Rashmi Gupta, Rami Albatal, and Duc-Tien Dang-Nguyen. 2017. Overview of NTCIR-13 Lifelog-2 Task. (2017), 6.

[15] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2021. Introduction to the Fourth Annual Lifelog Search Challenge, LSC'21. In *Proceedings of the 2021 International Conference on Multimedia Retrieval.* 690–691.

[16] Cathal Gurrin, Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Björn Þór Jónsson, Jakub Lokoč, Wolfgang Hürst, Minh-Triet Tran, and Klaus Schoeffmann. 2020. Introduction to the third annual lifelog search challenge (LSC'20). In *Proceedings of the 2020 International Conference on Multimedia Retrieval.* 584–585.

[17] Cathal Gurrin, Alan F Smeaton, and Aiden R Doherty. 2014. Lifelogging: Personal big data. *Foundations and trends in information retrieval* 8, 1 (2014), 1–125.

[18] Cathal Gurrin, Liting Zhou, Graham Healy, Bjorn Thor Jonsson, Duc Tien Dang Nguyen, Jakub Lokoc, Minh-Triet Tran, Wolfgang Hurst, Luca Rossetto, and Klaus Schoeffmann. 2022. An Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *ICMR '22, The 2022 International Conference on Multimedia Retrieval.* ACM, Newark, NJ, USA.

[19] Morgan Harvey, Marc Langheinrich, and Geoff Ward. 2016. Remembering through lifelogging: A survey of human memory augmentation. *Pervasive and Mobile Computing* 27 (2016), 14–26.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, Las Vegas, NV, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

[21] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *arXiv:1709.00029 [cs]* (Feb. 2019). http://arxiv.org/abs/1709.00029 arXiv: 1709.00029 version: 2.

[22] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoč, Andreas Leibetseder, František Mejzlík, Ladislav Peška, Luca Rossetto, et al. 2022. Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. *International Journal of Multimedia Information Retrieval* 11, 1 (2022), 1–18.

[23] Silvan Heller, Luca Rossetto, Loris Sauter, and Heiko Schuldt. 2022. Vitrivr at the Lifelog Search Challenge 2022. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22).* Association for Computing Machinery, New York, NY, USA, 27–31. https://doi.org/10.1145/3512729.3533003

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. *arXiv:2102.05918 [cs]* (June 2021). http://arxiv.org/abs/2102.05918 arXiv: 2102.05918.

[25] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2016. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. *arXiv:1612.06890 [cs]* (Dec. 2016). http://arxiv.org/abs/1612.06890 arXiv: 1612.06890.

[26] Pei-Wei Kao, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. ConvLogMiner: A Real-Time Conversational Lifelog Miner. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence.*

[27] 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* (Taipei, Taiwan) *(LSC '21).* Association for Computing Machinery, New York, NY, USA, 3–6. https://doi.org/10.1145/3463948.3469255

[28] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster).*

[29] Gregor Kovalčík, Vít Škrhak, Tomáš Souček, and Jakub Lokoč. 2020. VIRET Tool with Advanced Visual Browsing and Feedback. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge.* ACM, Dublin Ireland, 63–66. https://doi.org/10.1145/3379172.3391725

[30] Jiayu Li, Min Zhang, Weizhi Ma, Yiqun Liu, and Shaoping Ma. 2020. A Multi-level Interactive Lifelog Search Engine with User Feedback. In *Proceedings of the Third Annual Workshop on Lifelog Search Challenge.* ACM, Dublin Ireland, 29–35. https://doi.org/10.1145/3379172.3391720

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]* (July 2019). http://arxiv.org/abs/1907.11692 arXiv: 1907.11692.

[32] Jakub Lokoč, František Mejzlik, Patrik Veselý, and Tomáš Souček. 2021. Enhanced SOMHunter for Known-Item Search in Lifelog Data. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21).* Association for Computing Machinery, New York, NY, USA, 71–73. https://doi.org/10.1145/3463948.3469074

[33] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[34] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations.*

[35] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. 2018. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Transactions on Geoscience and Remote Sensing* 56, 4 (April 2018), 2183–2195. https://doi.org/10.1109/TGRS.2017.2776321 arXiv: 1712.07835 version: 1.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]* (Feb. 2021). http://arxiv.org/abs/2103.00020 arXiv: 2103.00020.

[37] Ricardo Ribiero, Alina Trifan, and Antonio J. R. Neves. 2022. MEMORIA: A Memory Enhancement and MOment RetrIeval Application for LSC 2022. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22).* Association for Computing Machinery, New York, NY, USA, 8–13. https://doi.org/10.1145/3512729.3533011

[38] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]* (Feb. 2020). http://arxiv.org/abs/1910.01108 arXiv: 1910.01108.

[39] Mohit Shah, Brian Mears, Chaitali Chakrabarti, and Andreas Spanias. 2012. Lifelogging: Archival and retrieval of continuously recorded audio using wearable devices. In *2012 IEEE International Conference on Emerging Signal Processing Applications.* 99–102.

[40] Florian Spiess and Heiko Schuldt. 2022. Multimodal Interactive Lifelog Retrieval with Vitrivr-VR. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22).* Association for Computing Machinery, New York, NY, USA, 38–42. https://doi.org/10.1145/3512729.3533008

[41] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks* 0 (2012), –. https://doi.org/10.1016/j.neunet.2012.02.016

[42] Ly-Duyen Tran, Thanh Cong Ho, Lan Anh Pham, Binh Nguyen, Cathal Gurrin, and Liting Zhou. 2022. LLQA - Lifelog Question Answering Dataset. In *MultiMedia Modeling.* Springer International Publishing, Cham, 217–228.

[43] Ly-Duyen Tran, Manh-Duy Nguyen, Binh Nguyen, Hyowon Lee, Liting Zhou, and Cathal Gurrin. 2022. E-Myscéal: Embedding-Based Interactive Lifelog Retrieval System for LSC'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22).* Association for Computing Machinery, New York, NY, USA, 32–37. https://doi.org/10.1145/3512729.3533012

[44] Ly-Duyen Tran, Manh-Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. 2021. Myscéal 2.0: a revised experimental interactive lifelog retrieval system for LSC'21. In *Proceedings of the 4th Annual on Lifelog Search Challenge.* 11–16.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems,* Vol. 30. Curran Associates, Inc.

[46] Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation Equivariant CNNs for Digital Pathology. *arXiv:1806.03962 [cs, stat]* (June 2018). http://arxiv.org/abs/1806.03962 arXiv: 1806.03962.

[47] Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. ActionCLIP: A New Paradigm for Video Action Recognition. *arXiv:2109.08472 [cs]* (Sept. 2021). http://arxiv.org/abs/2109.08472 arXiv: 2109.08472.

[48] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2021. Robust fine-tuning of zero-shot models. *CoRR* abs/2109.01903 (2021). arXiv:2109.01903 https://arxiv.org/abs/2109.01903

[49] Yang Yang, Hyowon Lee, and Cathal Gurrin. 2013. Visualizing lifelog data for different interaction platforms. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems.* 1785–1790.

[50] TengQi Ye. 2018. *Visual Object Detection from Lifelogs using Visual Non-lifelog Data.* Ph. D. Dissertation. Dublin City University.

[51] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. *arXiv:2111.03930 [cs]* (Nov. 2021). http://arxiv.org/abs/2111.03930 arXiv: 2111.03930.

[52] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. 2021. PointCLIP: Point Cloud Understanding by CLIP. *arXiv:2112.02413 [cs]* (Dec. 2021). http://arxiv.org/abs/2112.02413 arXiv: 2112.02413.

[53] Liting Zhou and Cathal Gurrin. 2022. Multimodal Embedding for Lifelog Retrieval. In *International Conference on Multimedia Modeling*. Springer, 416–427.